HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2016 May 21.

Published in final edited form as:

J Chem Inf Model. 2015 July 27; 55(7): 1483–1494. doi:10.1021/acs.jcim.5b00030.

Relating Essential Proteins To Drug Side-Effects Using Canonical Component Analysis: A Structure-Based Approach

Tianyun Liu and

Department of Genetics, Stanford University

Russ B Altman

Department of Genetics and Department of Bioengineering, Stanford University

Tianyun Liu: tianyunl@stanford.edu; Russ B Altman: russ.altman@stanford.edu

Abstract

The molecular mechanism of many drug side-effects is unknown and difficult to predict. Previous methods for explaining side effects have focused on known drug targets and their pathways. However, low affinity binding to proteins that are not usually considered drug targets may also drive side-effects. In order to assess these alternative targets, we used the 3D structures of 563 essential human proteins systematically to predict binding to 216 drugs. We first benchmarked our affinity predictions with available experimental data. We then combined singular value decomposition and canonical component analysis (SVD-CCA) to predict side-effects based on these novel target profiles. Our method predicts side-effects with good accuracy (average AUC: 0.82 for side effects present in < 50% of drug labels). We also noted that side-effect frequency is the most important feature for prediction, and can confound efforts at elucidating mechanism; our method allows us to remove the contribution of frequency and isolate novel biological signals. In particular, our analysis produces 2768 triplet associations between 50 essential proteins, 99 drugs and 77 side-effects. Although experimental validation is difficult because many of our essential proteins do not have validated assays, we nevertheless attempted to validate a subset of these associations using experimental assay data. Our focus on essential proteins allows us to find potential associations that would likely be missed if we used recognized drug targets. Our associations provide novel insights about the molecular mechanisms of drug side-effects, and highlight the need for expanded experimental efforts to investigate drug binding to proteins more broadly.

Correspondence to: Russ B Altman, russ.altman@stanford.edu.

ASSOICATED CONTENT

Supporting Information Available:

Figure S1 PF-affinity scores of celecoxib against 563 essential proteins; Figure S2 Performance of SVD-CCA on our datasets; Figure S3 Characteristics of side-effect data from Yamanish's publication; Figure S4 Performance of SVD-CCA on the dataset from Yamanishi's publication; Figure S5 Data sparsity of the attribute weights from SVD-CCA; Figure S6 Statistics of the novel associations identified by SVD-CCA; Figure S7 The predicted relationships of drug, protein and side-effects for NSAIDs and antidepressants; Table S1 The most promiscuous drugs; Table S2 The most promiscuous targets; Table S3 Statistics of the novel associations identified by SVD-CCA; Table S4 The validated pairs of drug and proteins in the 2768 novel relationships.

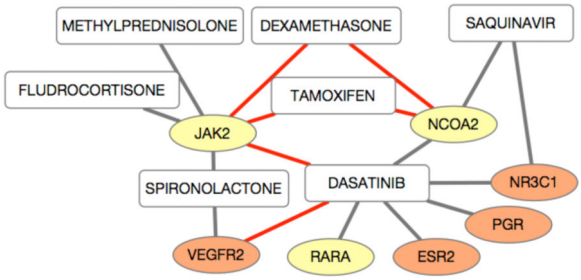
We also include addition information for the methods of SVD-CCA. The complete information of "Human Esseantial Proteins" and the predicted 2768 relationships (drug, gene, side-effects) are made available at http: http://simtk.org/SIDE_EFFECT.

This material is available free of charge via the Internet at http:///pubs.acs.org.

CONFILIT OF INTEREST

The authors declare no competing financial interest.

Graphical Abstract



INTRODUCTION

Drug discovery projects aim to develop highly selective compounds for a therapeutically relevant target while avoiding side-effects. The ability to predict side-effects is therefore valuable, especially if the underlying molecular pathways can be elucidated. Previous studies on side-effects have focused on using known drug targets and pathways as primary candidates to explain drug side-effects[1234567–8]. These studies implicitly assume a causal connection between the known targets and side-effects, while a drug's other binding activities (to proteins not considered targets or known to participate in side effects) are often ignored. This focus stems from an understandable desire to focus on targets and pathways that are known to produce drug-response phenotypes. However, recent literature suggests that low affinity binding to proteins that are not known drug targets and are not normally associated with drug response may also contribute to side-effects[9–10]. The hypothesis is that decreased selectivity for the desired target correlates with an increase of side-effect frequency resulting from unwanted binding to other proteins. Binding of drugs to their targets and to other unintended proteins, or "off-targets" may together explain the spectrum of efficacies and side-effects observed for many drugs.

To understand drug side-effects in a systematic and unbiased manner, we would ideally like a complete matrix of small molecule drugs and their binding affinities to all proteins. Such a data set would allow us to correlate global protein binding patterns to side-effect profiles. However, such large-scale binding assays are not generally available[1112]. Indeed, available biochemical data are typically biased towards known targets[13–14], and so we do not have complete information about the binding profile of small molecules to proteins that may be critical to cellular physiology, but which are not recognized drug targets. It is thus difficult to test ideas about unintended binding with experimental data because available datasets focus almost exclusively on known drug targets. We have surveyed the high confidence datasets from ChEMBL[15] and BindingDB[16] and found that there are on average 15 unique assays for each drug and two assays for each protein (unpublished data). The most influential research on revealing molecular mechanisms of side-effects are from Pfizer research (Biospectra)[17–19] and Novartis research[20], respectively. The original work of Biospectra by Fliri et al in 2005 is the first large-scale screening between 1045 drugs and 92 proteins. This works identified important molecular mechanisms of drug clinical effects and has been the foundation of many following research on drug side-effects[21222324]. The broad-scale *in vitro* pharmacology profiling by Novartis research analyzed drug promiscuity against 220 targets

(including 73 unintended targets)[20]. Unfortunately, most datasets do not sample the large set of proteins critical to cellular function but not considered drug targets, making it difficult to create an unbiased estimate of drug promiscuity. We thus turn to computational prediction of affinity to proteins not considered targets and estimation of drug promiscuity.

Computational prediction of drug-protein interactions offers an alternative to comprehensive experimental screens. These predictions usually employ methods in two categories. The first one estimates a drug's binding to a protein by considering the chemical properties of a drug such as its fingerprint or chemical structure[2526]. The other category explores the protein's small molecule binding profile and calculates similarity between two protein sites in order to predict new binding sites for existing drugs[272829]. Since most methods for predicting drug binding are structure-based, it is not possible to conduct a computational estimation over the entire human genome, because only 20% of human proteome is available as 3D structures (rough estimation, unpublished data).

Yu et al suggested that the number of essential genes that bind a drug, and not the number of known targets of the drug, is a primary determinant of side-effects[30]. Inspired by this observation, we assembled a representative set of 3D structures of "human essential proteins" identified by Silva et al[3130] using a large scale of RNAi screen. They identified 1830 essential genes important for the proliferation and survival of five cell lines derived from human mammary tissue. These genes target many important cellular metabolic and regulatory networks. We used the 3D structures for these essential genes to predict their binding to drugs, and subsequently used these predictions to estimate drug promiscuity and side-effect profiles.

To systematically relate drug characteristics (known targets, pathways, or other molecular basis) to side-effects, existing methods use machine learning algorithms and network-inference methods[732333, 65]. Given a set of drug characteristics, these algorithms can predict side-effects, but have not been shown to elucidate the molecular mechanisms for these side-effects. We consider the problem of mapping side-effects to drugs and proteins as a linear algebra problem. Let X be a matrix relating drugs to protein binding and Y be a matrix relating drugs to side-effects; the fundamental challenge is to relate specific attributes of X to those of Y. The canonical components analysis (CCA) is a method of finding linear relationships between two multidimensional variables[346, 35]. It finds basis vectors for X and Y such that the correlations between the projections of these matrices onto the basis vectors are maximized. We use this method to obtain a lower dimensional subspace that jointly associates drug binding (X) and drug side-effects (Y)[3536]. Given these basis vectors, new drugs can be projected onto this subspace to predict likely associated side-effects. Importantly, this process is transparent and allows us to determine the specific protein bindings that are associated with specific side effects. In the CCA optimization process, we adopt singular value decomposition (SVD) to obtain projected subspaces, such that the highest ranked subspace represents the dominant data feature. Therefore, we are able to extract the dominant associations or correlations between side-effects and predicted protein binding for individual drugs, providing insight into the mechanism of the side effects. Our ability to experimentally validate our findings is limited (Only about ~5% the proteins we have used in this work have high confidence experimental assays available in ChEMBL and

BindingDB). Our main purpose of this work is to present a novel method for relating essential proteins to side-effects, with prospective validation not being considered at the current stage. Nonetheless, our investigation of the limited experimental data that is available suggests that promiscuous binding across the proteome should be investigated experimentally and considered in the evaluation of drug side-effect profiles.

RESULTS

1. Create key datasets for our analysis

We mapped the 1830 human genes reported as essential from a large-scale RNAi screen in human mammary cells[31] to the PDB, resulting in 563 human essential proteins for which we have 3D structures, referred as the Essential Protein Dataset. They belong to 306 Pfam families and 198 superfamilies.

We assembled the Drug Binding Dataset by collecting 216 small molecule drugs, for which we have 3D structures co-crystallized with proteins (usually their targets) available in the PDB. We also retrieved their side-effect profiles from SIDER2[37], an online database containing information extracted from package inserts using text minding methods. The 978 binding sites for the 216 drugs represent 459 unique proteins, 298 Pfam and 164 superfamilies. There are only 39 superfamilies shared between the 164 superfamilies in the Drug Binding Dataset and the 198 superfamilies in the Essential Protein Dataset. Similarly, there are only 79 Pfam definitions shared between the 298 Pfam in the Drug Binding Dataset and the 306 Pfam in the Essential Protein Dataset, suggesting that the majority of proteins in the Essential Protein Dataset are not known to bind drugs—and thus are not typically considered targets.

We also constructed the IC50 Assay Dataset, which constitutes 234 high-confidences experimental assays (IC50) derived from ChEMBL and BindingDB between 47 unique drugs and the 94 unique human essential proteins.

2. Validate the use of PocketFEATURE to predict affinities

We employed our previously reported method, PocketFEATURE[29], to predict affinity scores (PF-affinity) between the 216 drugs in Drug Binding Dataset and the 563 proteins in the Essential Protein Dataset, resulting in 121,608 scores between drugs and proteins. Conceptually, each drug is represented as a vector of 563 predicted affinities.

For aggregated analysis of PF affinity scores, the 234 drug-protein pairs were divided into three groups according to their predicted affinity scores. Figure 1 shows the IC50 values of each group in boxplots. For the group of drug-protein binding that has PF-affinity scores lower than −4.0, the average IC50 value is only 865nM. As the PF-affinity score increases, the average IC50 increases accordingly. The bottom panel shows the histogram of IC50 values in different affinity score groups. In the group with affinity scores lower than −4.0, most assays have IC50 values lower than 10uM. This figure suggests that the PocketFEATURE affinity scores are a rough estimate of the potential binding of a drug to an essential protein. As an example, Figure S1 shows that the affinity scores between celecoxib

and six proteins correlate with the corresponding log (IC50) values that are available in the IC50 Assay Dataset.

For each of the 216 drugs, we estimated its potential for binding the 563 essential proteins by calculating the PF-affinity scores. Using a score cutoff of −2.0, we counted the predicted number of essential proteins bound for a given drug, as an estimate of the drug's promiscuity. Figure 2A shows the histogram of the number of the predicted essential proteins bound (score cutoff −2.0) for each of the 216 drugs. The most promiscuous drugs (with number of binding to essential proteins >200) are: troglitazone, dasatinib, sorafeib, aliskiren, imatinib and nilotinib (Supporting Information Table S1). About 15% drugs (34 of 216) bind to 20% or more of the 563 human essential proteins. More than 30% drugs bind to more than 10% of the essential proteins. This is consistent with the in house safety screen by Novartis research, which have reported that more than 20% drugs were found to bind to 10–20% of the profiling targets with an IC50 lower than 5uM. Figure 2B shows the histogram of the predicted number of essential proteins bound (score cutoff −2.0) for each of the 563 human essential proteins. About 70% of essential proteins are predicted to bind to less than 30 drugs and are classified as low promiscuity. The most promiscuous proteins are: B-cell CLL/lymphoma 2 (BCL2), E1A binding protein p300(EP300), Janus kinase 2(JAK2), adenylate kinase 3-like 2 (AK4), coagulation factor II (thrombin) receptor (F2R), nuclear receptor subfamily 5 (NR5A2) (Table S2).

3. Estimate side-effects using the predicted bindings

Figure 3 shows the histogram of the percentage of drugs (of a total 216) associated with each side-effect. There are 24 side-effects that are observed in more than 50% of the 216 drugs, including fatigue, heartburn, erythema, constipation, insomnia, anorexia, thrombocytopenia, fever, edema, anemia, hypersensitivity, dizziness, colic, diarrhea, rash, vomiting, nausea. We removed these 24 side-effects, creating a subset (SE-subset-50) that contains 1276 side-effects. To study side-effects that are even less common, we further constructed a subset containing 1115 side-effects that are observed only in 20% or fewer of the 216 drugs (SE-subset-20, see Method).

We grouped the 216 drugs according to the number of observed side-effects. The average number of side-effects in each group correlate with the average number of the predicted essential proteins bound, with a correlation coefficient of 0.88 (p-value < 0.02, Figure 4).

We used the SVD-CCA process to compute the linear relationship between the predicted affinity to human essential proteins (X) and the observed side-effects (Y) by projecting X and Y into multiple subspaces. Given a query drug with predicted affinities (X), we then estimated its side-effects (Y). The AUC values for side-effect predictions in the leave-one-out cross validation is 0.82 and 0.73 for SE-subset-50 and SE-subset-20, respectively (Figure 5 top panel and Supporting Information Figure S2). SVD-CCA achieves its best performance when using ten sets of subspaces (termed "canonical component", or CC). For predicting side effects, SVD-CCA performs better when it characterizes each drug based on the predicted essential proteins bound (for the X matrix), compared to characterizing the drug based on its known binding to recognized drug targets (Figure 5 bottom panel). Unlike the SVD-CCA performance on predicted essential proteins bound, the binding to known

drug targets yields only one informative CC, and its performance is not improved by adding additional CCs to predict side-effects.

In addition, SVD-CCA outperforms the traditional CCA algorithm, in terms of predicting side-effects. The average AUC of SVD-CCA is 0.82 on subset-50 and that of traditional CCA is 0.7 (Method SVD-CCA process: two options for solving CCA).

4. Separate frequency contribution from biological influences

The performance of our side-effect prediction is improved as we increase the number of canonical components used (Figure 5 top panel). Figure 6A illustrates the most informative canonical component CC#1; the attribute weights assigned to each side-effect correlate very well with side-effect frequency. However, there is no clear correlation between side-effect frequency and the weights extracted in the other components (CC#2 through CC#10) (Figure 6B). We conclude that these components must contain information about the biological association of side-effects with essential proteins. We thus sought to extract these relationships from the data.

5. Extract novel biological contributions within the data

In order to analyze the information contained in CC#2 through CC#10, we evaluated the attribute weights with extreme values. Figure 7 compares the weights in each canonical component extracted from SE-subset-20 to those from random permutation tests. Each CC contains two weight vectors: alpha provides weights for the human essential proteins and beta provides weights for side-effects. It shows that the extreme values in CC#2, CC#3, CC#4, CC#5 and CC#6 are unlikely to occur by chance and represent statistically significant signals ($p <= 0.01$, with 100 permutation tests). These signals relate particular essential proteins to particular side-effects, providing insight into the mechanism of these side-effects. We do not observe extreme weights in CC#7 and subsequent CCs, suggesting that the biological signal fades after CC#6.

6. Associate drug, protein and side-effect

Using the extreme associations observed in CC#2, CC#3, CC#4, CC#5 and CC#6, we identified 2768 triplets of [*drug, essential protein, side-effect*]. These sets involve 99 drugs, 50 essential proteins and 77 side-effects, resulting in 826 unique pairs of *essential proteins: side-effects* (Supporting Information Table S3). A total of 23 pairs of *drug: essential protein* are found in IC50 Assay Dataset, for which experimental IC50s are available. The 23 pairs involve a total of 120 drug: *essential protein*: *side-effect* relationships (Supporting Information Table S4). We also sought experimental evidence supporting the predicted associations between essential proteins and side-effects. We provide two examples here.

The first is the side-effect observed in CC#4: menstrual irregularities/disorders. Figure 8A shows the seven predicted essential proteins associated with menstrual irregularities: glucocorticoid receptor (NR3C1), glucocorticoid Nuclear Receptor 2(NCOA2), progesterone-receptor (PGR), retinoic-acid-receptor (RARA), estrogen-receptor (ESR2), Janus kinase 2 (JAK2), and vascular endothelial growth factor 2(VEGFR2). Experimental assays in our IC50 Assay Dataset confirmed six pairs of bindings between essential proteins

and drugs (red edge): tamoxifen to NCOA2 and JAK2, dexamethasone to NCOA2 and JAK2 and dasatinib to JAK2 and VEGFR2. We also inspected the newly discovered associations reported by Novartis[20], which revealed ten proteins responsible for menstrual irregularities. Four of our seven predicted essential proteins overlap with these ten: estrogen-receptor (ESR2), progesterone-receptor (PGR), glucocorticoid receptor (NR3C1) and vascular endothelial growth factor 2 (VEGFR2). The other six include three hormone receptors or hormone binding proteins, and three kinases.

The second example is observed in CC#3: hypocalcemia, abnormally low blood calcium levels. Figure 8B shows the seven predicted associations: apo-lipoprotein (APOD), DNA polymerase beta (POLB), histone deacetylase 8(HDAC8), hypoxia-inducible factor 1-alpha inhibitor (FIH1), glyoxalase I (GLO1), matrix metalloproteinase-3 (MMP3), matrix metalloproteinase-7 (MMP7), peroxisome proliferator-activated receptor alpha (PPARA), sex steroid-binding protein (SBP) and vitamin D3 receptor (VDR). We also found published evidence that two proteins are associated with hypocalcemia: calcium-sensing receptor (CASR) and VDR[38]. Naturally occurring mutations in CASR cause hypocalcaemia or hypercalcaemia. The 3D structure of CASR is not available so we were not able to estimate its probability of binding to hypocalcemia related drugs. In knockout mice, genetic inactivation of VDR leads to hypocalcemia[38]. We have predicted VDR as an important interaction that may interact with saquinavir (HIV drug) and three cancer drugs. In addition, ganciclovir is known to interact with POLB[39] and saquinavir binds to PPARA[40]. Since hypocalcemia is associated with a variety of drugs, the predicted molecular networks provide insights to the mechanism of hypocalcemia.

7. Analyze novel predictions

From the associations observed in CC#2, CC#3, CC#4, CC#5 and CC#6, we identified 2768 triplets of [*drug, essential protein, side-effect*]. Most of these associations are novel. They are made available at https://simtk.org/home/side-effect/. The 2768 associations involve 50 essential proteins. These proteins have predicted binding to drugs ranging from 2 to 33 (Supporting Information Figure S6), considered as low promiscuity proteins (Figure 2B). The molecular functions of these 50 essential proteins are shown in Figure 9. Essential proteins classified as nuclear receptors (NR))(GO:0004879), DNA-binding transcription factors (GO:0003700) and protein-binding transcription factors (GO:0000988) are enriched. The most frequently observed proteins in the 2768 joint sets are: glucocorticord receptor (complexed with nuclear receptor coactivator 2), progesterone receptor, transcriptional intermediary factor2 and retinonic acid receptor RXR-alpha (Supporting Information Table S3).

Another view of these novel predictions is to group them by drugs. Figure S7-A shows the associations for five non-steroidal anti-inflammatory (NSAID) drug: celecoxib, valdecoxib, fenoprofen, naproxen and ibuprofen. These drugs cause 38 *significant side-effects* (see definition in Method section 5). However, only eleven of these side-effects are shared by two or more NSAIDs. Twelve side-effects are unique to celecoxib; six to valdecoxib, five to naproxen, three to ibuprofen and one to fenoprofen. The identified *significant essential proteins* are often unique to one NSAID, with only PPARG shared by celecoxib and

fenoprofen, and HIF1AN (transcription factor) shared by fenoprofen and ibuprofen. Figure S7-B shows the relationships extracted for six antidepressants: imipramine, sertraline, desipramine, amitriptyline, clomipramine and fluoxetine. They cause 51 side-effects, of which 19 are unique to fluoxetine only. Another 24 side-effects are shared by two or more antidepressants. Of the twelve *significant essential proteins*, seven are shared by two or more drugs.

DISCUSSION

1. Essential protein interactions contribute to side-effects

Current experimental assay data are not sufficient for identifying novel molecular mechanisms underlying side-effects, because they focus on known targets and do not consider binding to non-target proteins may drive side-effects. Moreover, current estimates of drug promiscuity are often based on assays that target known receptor families: G-protein coupled receptors (GPCRs), nuclear receptors, transporters, enzymes, and ion channels with known side-effect associations[41–4214]. To identify novel protein interactions contributing to side-effects, we have employed computational methods that estimate a more comprehensive set of protein binding affinities, selecting proteins that are not generally targets but are involved in key biological processes.

Our computationally predicted affinities are no doubt imperfect, but seem to be sufficiently precise for the purposes of this preliminary investigation of drug promiscuity, and the role of proteins not typically associated with drug response. The computationally predicted affinities show good correlation with experimental assays, suggesting that our predicted affinities are reasonable proxies for these measurements. Based on the predicted binding affinity (Figure 1), we have quantified drug promiscuity and found that about 15–30% drugs bind to 10–20% of the 563 human essential proteins (Figure 2). This is consistent with Novartis in-house safety screen data[414120]: more than 20% of all ligands were found to bind to 10–20% of the profiling targets (7–14 in absolute numbers) with an IC50 lower than 5uM. These data provide some confidence in the reliability of our computational profiling.

We observe that the number of essential proteins bound by a drug correlates with its observed number of side-effects. We suggest that these novel protein-drug interactions provide information about the molecular mechanism of the side-effects. Our SVD-CCA protocol produces 2768 specific associations between drugs, essential proteins, and side-effects, some of which are confirmed by existing experimental evidence. The predicted affinity scores are used in the SVD-CCA process for building associations between proteins and side-effects. Hence, drug protein interactions, or binding affinities play a key role in deciding such associations.

Previous research has understandably focused on known drug targets to explain molecular mechanisms of side-effects[1, 4, 6, 20]. According to[1], GPCRs contributed most to observed side-effects, compared to the other four important families of drug targets (nuclear receptor, ion channel, enzyme and any targets)[42]. Our results suggest that previous target-focused studies may have missed other important interactions, as suggested by Figure 8. The 2768 associations involve 520 unique pairs of drugs and essential proteins. Only four of these

pairs are known drug target pairs found in DrugBank[43], suggesting that our study scope discovers novel interactions. For example, the non-redundant set of 3D structures of essential proteins only contain three GPCRs due to limitations of structure-based methods. In our prediction, one of the three GPCRs are significantly associated with side-effects.

Of the 50 essential proteins that involve in the 2768 specific associations, nuclear receptors and transcription factors are enriched. (Figure 9 and Supporting Information Table S3). The top three *essential proteins* associated with most side-effects are glucocorticoid receptor (NR3C1), glucocorticoid receptor 2 (NCOA2) and progesterone receptor (PGR), highlighting the importance of hormone regulation as a source of side-effects. The next group of frequently observed essential proteins are involved in DNA replication and transcription, including DNA polymerase beta (POLB), factor inhibiting HIF-1 (FIH1), hypoxia inducible factor-1 (HIF1A), and spindling-1 (SPIN1). These suggest a set of side-effects that result from disruption of cell division and expression regulation. These proteins are not recognized drug targets, and illustrate the utility of analyzing essential proteins. The 50 most significant essential proteins bind between 2 and 33 drugs and are central to critical cellular pathways, suggesting that many side-effects stem from disruption of these more fundamental pathways and not necessarily interaction with known drug targets. Our predictions suggest association between genes and side-effects. Drug binding to these genes could be the driving force of downstream pathways and cellular responses that lead to side-effects. These pathways and cellular responses are not studies in this work. Other important aspects of drug activities (absorption, distribution, metabolism and elimination) should be investigated in our future work. In addition, we do not have the ability to distinguish the potential roles of protein homologs in the mechanism of side-effects. We have used predicted essential proteins, but their pockets are likely similar to the pockets of their close homologs (paralogs), and so some side-effects may be mediated through the homologs.

It is not surprising that there are relatively few available assays with which to compare our predictions--the Essential Protein Dataset generally has not been screened for drug interactions. Of the 121,608 pairs of drugs and essential proteins for which we predicted binding, only 234 pairs (0.2%) that have been tested experimentally. For our 520 pairs of high confidence interactions between drugs and essential proteins, there are 23 pairs (4.5%) that occur in the IC50 Assay Dataset.

2. SCD-CCA provides strategies that correlate high dimension variables and decouple biological mechanism from the frequency effects

We discovered novel associations between specific proteins and side-effects, and did not focus on predicting side-effects alone. We chose the CCA algorithm because it provides a method for uncovering the relationship between two variables by projecting them into a joint subspace[3536]. It computes a canonical component that provides weights for key attributes (Within one canonical component, high weighted attributes from two variables are associated.). In our work, the two variables represent the drugs' side-effects and their binding to essential proteins. Our analysis is predicated on the assumption that these variables should be highly correlated in some projected subspaces, because they have a causal relationship. Intuitively, our objective is to find k sets of projection weights (canonical

components) that project input matrices onto subspaces in which the correlations between projected vectors (corresponding to a particular drug) are maximized. We then extract specific associations between side-effects and their binding to essential proteins in each subspace.

Our results highlight two critical issues relevant to predicting side-effects. First, they are often biased because they focus on the analysis of known drug targets. Second, they often confound side-effect frequency with biological mechanisms. The success at predicting side-effects often depends on the dominant effect of frequency (Supporting Information Figure S4). When benchmarking the performance of SVD-CCA on the published dataset by Yamanishi[342], we have found that using only the first canonical component alone can achieve excellent performance (AUC=0.92). Figure S4B further shows that the information extracted in the first canonical component reflects side-effect frequency. (*Their dataset contains data of known targets for 674 drugs and their observed side-effects that includes high frequently observed side-effects. Predicting side-effects that includes frequently observed side-effects tend to result in high performances*.) Other studies that predict side-effects using models built based on side-effect data from SIDER using other machine learning methods also implicitly rely on side-effect frequencies[6322]. Unfortunately, these approaches generally do not decouple frequency from the other pertinent biological signals.

Our combined SVD-CCA algorithm is able to decouple the side-effect frequency from other biological factors. As shown in Figure 5 (top panel), when the inputs are the predicted essential proteins bound, the performance of predicting side-effects improves as the number of CCs increases. Importantly, other highly ranked canonical components contain biological information for understanding side-effects. Indeed, we used the SVD-CCA analysis to separate specific biological associations from the dominant side-effect frequency signal. In the contrast, when using binding to recognized drug targets as input, the performance is not improved as the number of CCs used in predicting side-effects increases (Figure 5 bottom panel). This suggests that components other than the first component extracted do not have useful information for predicting side-effects. It is not surprising given the low rank of the drug known target data (lack of information). This is also reflected in our benchmark on the published dataset by Yamanishi[432], which also uses recognized drug targets as input for predicting side-effects. Figure S4A shows that the performance is **not** improved as the number of CCs increases. In summary, with the ability of decoupling frequency from biological factors, SVD-CCA has advantages of extracting useful biological factors when the input molecular data is informative.

3. Canonical components representing specific characters in different subspaces

We obtained biological signals for side-effects by analyzing canonical components other than the first (which encodes frequency). We see strong signals for essential proteins that are involved in some side-effects, and we have shown novel associations between essential proteins, drugs for menstrual irregularities (observed in canonical component CC#4) and hypocalcemia (observed in canonical component CC#3). Direct evidence for protein and side-effect relationships is rare; hence predicted associations are useful to understand the molecular mechanisms and generating testable hypotheses.

We have linked menstrual irregularities and hypocalcemia both to essential proteins that are predicted to bind saquinavir and dasatinib. However, these two side-effects are associated with non-overlapping sets of essential proteins. For menstrual irregularities, many of the essential proteins are related to hormone binding and hormone receptors (CC#4). For hypocalcemia, the essential proteins are for translation regulation (CC#3). Thus, the binding profiles of saquinavir and dasatanib can be dissected and associated with specific molecular pathways that cause different side effects. This observation highlights the ability of SVD-CCA to decouple the associations of essential proteins and assign them to the side-effect information in different subspaces[36]. That is, given that a particular subset essential proteins are involved in a side-effect, SVD-CCA discovers this the projection in which this subset is correlated with the side-effect data.

METHODS

1. Datasets

Drug Binding Dataset collects 216 small molecule drugs that satisfy the following standards. (1) The 3D conformations of a drug's binding sites are in PDB; (2) The side-effect records of the drug is in SIDER2, an online database containing drug side-effect associations extracted from package inserts using text minding methods[37]. Drug Binding Dataset contains 978 binding sites for the 216 drugs, representing 586 KEGG pathways, 298 Pfam and 164 superfamilies.

Essential Protein Dataset collects 563 proteins that satisfy the following standards. (1) A biological meaningful small molecule ligand[16] is known to bind to the protein and the 3D conformations of the binding site are available in PDB; (2) The protein belongs to the 1830 human proteins reported as essential from a large-scare RNAi screen in human mammary cells[31]. This dataset represents 205 unique KEGG pathways, 306 Pfam families and 198 superfamilies.

IC50 Assay Dataset represents a total of 234 assays (IC50) between 47 unique drugs and 94 unique human essential proteins (https://simtk.org/home/side-effect). These set of high confidence assays are derived from 2385 unique assays (IC50) between 166 small molecule drugs and 1060 proteins from BindingDB and ChEMBL that satisfy the following standards. (1) At least one IC50 value is recorded for the assay. (2) For duplicated assays between one pair of protein and drug, a best-reported IC50 is used. (3) For an assay from ChEMBL, its confidence level has to be nine or above. (4) The drug belongs to our Drug Binding Dataset. Of the 563 human essential proteins, only 94 unique proteins can be found and lead to the 234 assays.

To derive side-effect subsets, we first construct a dataset that satisfy two conditions. (1) The drug belongs to our Drug Binding Dataset. (2) Side-effects are observed in three or more of the 216 drugs in Drug Binding Dataset. This leads to a dataset of 216 drugs and 1300 side-effects. We then derive two subsets: SE-subset-50 contains side-effects that are observed in more than 50% of the 216 drugs; SE-subset-20 contains are side-effects that observed only in 20% or less of the 216 drugs.

2. Predict affinities

We employ a previously developed method PocketFEATURE[29], which compares similarities between two binding sites, to calculate the probability of binding for a given pair of a drug and a target protein.

In our study, a drug is actually represented by its binding site properties[44], which are defined by protein residues within 6 Angstroms of the drug molecule. A target protein is represented by the binding site of the largest biological meaningful molecule co-crystalized with the protein.

A defined site (a set of residues) is then described with the physiochemical and structural environments surrounded around each residue. For each residue in a site, we choose a central functional atom and calculate the FEATURE microenvironment around the center. Specifically, FEATURE system calculates a set of 80 physicochemical properties collected over six concentric spherical shells (total 480 properties = 80 properties×6 shells) centered on the predefined functional center. FEATURE microenvironment refers to the local, spherical region in the protein structure that may encompass residues discontinuous in sequence and structure. PocketFEATURE calculates site similarities by matching microenvironments between two sites. A complete description of FEATURE and PocketFEATURE can be found in[29, 44–45].

The similarity between a drug's binding site and a potential site in a target protein estimated by PocketFEATURE is referred as "PF-affinity score" in this study. A more negative score suggests a higher probability of the drug's binding to the target protein. A cutoff of −2.0 is usually used to define if a drug binds to a protein.

3. SVD-CCA process

We construct two matrices: X is the affinity scores between the 216 drugs and ***p*** proteins; Y is the observation of ***q*** side-effects in the 216 drugs (from either SE-subset-50, SE-subset-20, or control subsets). Traditional canonical correlation analysis is able to correlate linear relationship between X and Y. The event is that drugs' binding to essential proteins (X) causes side-effects (Y) and this event (X=>Y). CCA seeks for the weights for X and Y (α and β, respectively), such that X and Y are transformed back to the mostly likely causal relationship, where $\mathrm{corr}\left(\alpha^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}, \beta^{\mathrm{T}}\mathrm{Y}^{\mathrm{T}}\right) = \frac{\alpha^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{Y}\beta}{|\alpha^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}||\beta^{\mathrm{T}}\mathrm{Y}^{\mathrm{T}}|}$ is maximized. Here $\alpha = (\alpha_1, \ldots \alpha_p)^T$ and $\beta = (\beta_1, \ldots \beta_q)^T$.

Step 1. Estimate α β—We discuss two options for estimating $\underline{\alpha\ \beta}$.

The goal of the traditional CCA is to find weight α and β that maximize the following canonical correlations coefficient: $\mathrm{corr}\left(\alpha^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}, \beta^{\mathrm{T}}\mathrm{Y}^{\mathrm{T}}\right) = \frac{\alpha^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{Y}\beta}{|\alpha^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}||\beta^{\mathrm{T}}\mathrm{Y}^{\mathrm{T}}|}$. This can be solved by Lagrangian[35] that maximizes $\alpha^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{Y}\beta$ with the constraints $|\alpha^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}| = 1$ and $|\beta^{\mathrm{T}}\mathrm{Y}^{\mathrm{T}}| = 1$. This leads to the subspace with the maximized $\alpha^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{Y}\beta$. Then we can also obtain other subspaces by rankings of $\alpha^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{Y}\beta$. Here each pair of $\alpha^T X^T$ and $Y\beta$ in the corresponding

subspace is called a *canonical component*. That is, we are able to obtain multiple canonical components (subspaces), with multiple sets of α and β.

According to[36], the covariance matrix (X^TX *and* Y^TY) can be treated as a diagonal matrix. Therefore we maximize $\alpha^T X^T Y \beta$ with the constraints $|\alpha| = 1$ and $|\beta| = 1$.

To find α β, we have singular value decomposition (SVD) of X^TY

$$X^TY = U\Sigma V^T$$

where the orthogonal matrix U is defined by $U = [u_1, .., u_{10}, \ldots u_{100}, \ldots]$; $U^TU = I$ and I is an identity matrix where Σ is a diagonal matrix defined by $\Sigma = \begin{bmatrix} S_1 & & \\ & \ldots & \\ & & S_N \end{bmatrix}$ where the orthogonal matrix V is defined by $V = [v_1, .., v_{10}, \ldots v_{100}, \ldots]$; $V^TV = I$; Then we have $\alpha^T X^T Y \beta = \alpha^T U \Sigma V^T \beta = [\alpha^T u_1, .., \alpha^T u_N] \begin{bmatrix} S_1 & & \\ & \ldots & \\ & & S_N \end{bmatrix} \begin{bmatrix} v_1^T\beta \\ \ldots \\ v_N^T\beta \end{bmatrix}$ The maximum value of $\alpha^T X^T Y \beta$ is S_1 (the largest singular value) when $\alpha = u_1$ *and* $\beta = v_1$ (See Expanded View). Each pair of $\alpha^T X^T$ and $Y\beta$ is called a *canonical component*. Multiple components can be obtained by rankings of singular values.

We compare the two options by assessing their performance in predicting side-effects (Y) given a query drug (X).

Step 2. Calculate side-effect (y) of a given drug's binding profile (x)—In this step, we can select weights from multiple canonical components to calculate y[6]. We have tested the number of sets to achieve best performance in cross-validation. Ten sets of canonical components are used in this study to achieve optimal performance. This can be considered "dimension reduction". Note that the inverse below is pseudo-inverse.

$$B = [v_1, \ldots v_{10}]$$

$$A = [u_1, \ldots, u_{10}]$$

$$y = [BB^T]^{-1} BA^T x$$

We conduct a leave-one-out cross validation on the 216 drugs for benchmarking the performance.

4. Permutation tests

The permutation function in MatLab rearranges the dimensions of X, so that they are in the order specified by the vector order. The output X' has the same values of X but the order of

the subscripts needed to access any particular element is rearranged as specified by order. All the elements of order must be unique. For each X' and Y, we applied SVD-CCA and calculate AUC values from cross-validations.

From SVD-CCA calculation, we are able to extract multiple sets of weights (α *and* β). We plot α *and* β against to those extracted from permutation tests to exam the significance of extreme values of α *and* β.

5. Extract relationships between drug, essential protein and side-effect

We compare the CCs extracted from SE-subset-20 with those from the permutation tests. Each CC contains two vectors: alpha as weights of human essential proteins and beta as weights of side-effects. We use a weight absolute value cutoff 0.075 to collect essential protein and side-effects with extreme weights, named as "*significant essential protein*" and "*significant side-effect*", respectively. From CC#2, CC#3, CC#4, CC#5, CC#6, we extract 118 *significant essential proteins* and 86 *significant side-effects*. We track down the affinity scores of these proteins to the 216 drugs. Using a score cutoff of −2.5 (a more strict cutoff), pairs of drug: *significant essential proteins* are identified. When a drug is observed with a *significant side-effect*, the pair of drug: *significant side-effect* is also listed for further analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH GM102365, GM072970 and U54 HL117798. We thank A Gottlieb, S Rensi and Y Li for helpful discussion.

REFERENCE

1. Kuhn M, Al Banchaabouchi M, Campillos M, Jensen LJ, Gross C, Gavin AC, Bork P. Systematic Identification of Proteins That Elicit Drug Side Effects. Mol Syst Biol. 2013; 9:663. [PubMed: 23632385]
2. Yamanishi Y, Pauwels E, Kotera M. Drug Side-Effect Prediction Based on the Integration of Chemical and Biological Spaces. J Chem Inf Model. 2012; 52:3284–3292. [PubMed: 23157436]
3. Pauwels E, Stoven V, Yamanishi Y. Predicting Drug Side-Effect Profiles: A Chemical Fragment-Based Approach. BMC Bioinformatics. 2011; 12:169. [PubMed: 21586169]
4. Mizutani S, Pauwels E, Stoven V, Goto S, Yamanishi Y. Relating Drug-Protein Interaction Network with Drug Side Effects. Bioinformatics. 2012; 28:i522–i528. [PubMed: 22962476]
5. Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen XW, Matheny ME, Xu H. Large-Scale Prediction of Adverse Drug Reactions Using Chemical, Biological, and Phenotypic Properties of Drugs. J Am Med Inform Assoc. 2012; 19:e28–e35. [PubMed: 22718037]
6. Atias N, Sharan R. An Algorithmic Framework for Predicting Side Effects of Drugs. J Comput Biol. 2011; 18:207–218. [PubMed: 21385029]
7. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug Target Identification Using Side-Effect Similarity. Science. 2008; 321:263–266. [PubMed: 18621671]
8. Bauer-Mehren A, van Mullingen EM, Avillach P, Carrascosa Mdel C, Garcia-Serna R, Pinero J, Singh B, Lopes P, Oliveira JL, Diallo G, Ahlberg Helgee E, Boyer S, Mestres J, Sanz F, Kors JA,

Furlong LI. Automatic Filtering and Substantiation of Drug Safety Signals. PLoS Comput Biol. 2012; 8:e1002457. [PubMed: 22496632]

9. Chang RL, Xie L, Xie L, Bourne PE, Palsson BO. Drug Off-Target Effects Predicted Using Structural Analysis in the Context of a Metabolic Network Model. PLoS Comput Biol. 2010; 6:e1000938. [PubMed: 20957118]
10. Xie L, Xie L, Bourne PE. Structure-Based Systems Biology for Analyzing Off-Target Binding. Curr Opin Struct Biol. 2011; 21:189–199. [PubMed: 21292475]
11. Faller B, Wang J, Zimmerlin A, Bell L, Hamon J, Whitebread S, Azzaoui K, Bojanic D, Urban L. High-Throughput in Vitro Profiling Assays: Lessons Learnt from Experiences at Novartis. Expert Opin Drug Metab Toxicol. 2006; 2:823–833. [PubMed: 17125403]
12. Whitebread S, Hamon J, Bojanic D, Urban L. Keynote Review: In Vitro Safety Pharmacology Profiling: An Essential Tool for Successful Drug Development. Drug Discov Today. 2005; 10:1421–1433. [PubMed: 16243262]
13. Rask-Andersen M, Almen MS, Schioth HB. Trends in the Exploitation of Novel Drug Targets. Nat Rev Drug Discov. 2011; 10:579–590. [PubMed: 21804595]
14. Imming P, Sinning C, Meyer A. Drugs, Their Targets and the Nature and Number of Drug Targets. Nat Rev Drug Discov. 2006; 5:821–834. [PubMed: 17016423]
15. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. Chembl: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012; 40:D1100–D1107. [PubMed: 21948594]
16. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. Bindingdb: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. Nucleic Acids Res. 2007; 35:D198–D201. [PubMed: 17145705]
17. Fliri AF, Loging WT, Thadeio PF, Volkmann RA. Analysis of Drug-Induced Effect Patterns to Link Structure and Side Effects of Medicines. Nat Chem Biol. 2005; 1:389–397. [PubMed: 16370374]
18. Fliri AF, Loging WT, Thadeio PF, Volkmann RA. Biospectra Analysis: Model Proteome Characterizations for Linking Molecular Structure and Biological Response. J Med Chem. 2005; 48:6918–6925. [PubMed: 16250650]
19. Fliri AF, Loging WT, Thadeio PF, Volkmann RA. Biological Spectra Analysis: Linking Biological Activity Profiles to Molecular Structure. Proc Natl Acad Sci U S A. 2005; 102:261–266. [PubMed: 15625110]
20. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Cote S, Shoichet BK, Urban L. Large-Scale Prediction and Testing of Drug Activity on Side-Effect Targets. Nature. 2012; 486:361–367. [PubMed: 22722194]
21. Houck KA, Dix DJ, Judson RS, Kavlock RJ, Yang J, Berg EL. Profiling Bioactivity of the Toxcast Chemical Library Using Biomap Primary Human Cell Systems. J Biomol Screen. 2009; 14:1054–1066. [PubMed: 19773588]
22. Vegner L, Peragovics A, Tombor L, Jelinek B, Czobor P, Bender A, Simon Z, Malnasi-Csizmadia A. Experimental Confirmation of New Drug-Target Interactions Predicted by Drug Profile Matching. J Med Chem. 2013; 56:8377–8388. [PubMed: 24088053]
23. Parasuraman S. Prediction of Activity Spectra for Substances. J Pharmacol Pharmacother. 2011; 2:52–53. [PubMed: 21701651]
24. DiMaggio PA Jr, Subramani A, Judson RS, Floudas CA. A Novel Framework for Predicting in Vivo Toxicities from in Vitro Data Using Optimal Methods for Dense and Sparse Matrix Reordering and Logistic Regression. Toxicol Sci. 2010; 118:251–265. [PubMed: 20702588]
25. Adams JC, Keiser MJ, Basuino L, Chambers HF, Lee DS, Wiest OG, Babbitt PC. A Mapping of Drug Space from the Viewpoint of Small Molecule Metabolism. PLoS Comput Biol. 2009; 5:e1000474. [PubMed: 19701464]
26. Yera ER, Cleves AE, Jain AN. Chemical Structural Novelty: On-Targets and Off-Targets. J Med Chem. 2011; 54:6771–6785. [PubMed: 21916467]
27. Xie L, Bourne PE. Detecting Evolutionary Relationships across Existing Fold Space, Using Sequence Order-Independent Profile-Profile Alignments. Proc Natl Acad Sci U S A. 2008; 105:5441–5446. [PubMed: 18385384]

28. Xie L, Bourne PE. A Unified Statistical Model to Support Local Sequence Order Independent Similarity Searching for Ligand-Binding Sites and Its Application to Genome-Based Drug Discovery. Bioinformatics. 2009; 25:i305–i312. [PubMed: 19478004]
29. Liu T, Altman RB. Using Multiple Microenvironments to Find Similar Ligand-Binding Sites: Application to Kinase Inhibitor Binding. PLoS Comput Biol. 2011; 7:e1002326. [PubMed: 22219723]
30. Wang X, Thijssen B, Yu H. Target Essentiality and Centrality Characterize Drug Side Effects. PLoS Comput Biol. 2013; 9:e1003119. [PubMed: 23874169]
31. Silva JM, Marran K, Parker JS, Silva J, Golding M, Schlabach MR, Elledge SJ, Hannon GJ, Chang K. Profiling Essential Genes in Human Mammary Cells by Multiplex Rnai Screening. Science. 2008; 319:617–620. [PubMed: 18239125]
32. Yang L, Chen J, He L. Harvesting Candidate Genes Responsible for Serious Adverse Drug Reactions from a Chemical-Protein Interactome. PLoS Comput Biol. 2009; 5:e1000441. [PubMed: 19629158]
33. Wallach I, Jaitly N, Lilien R. A Structure-Based Approach for Mapping Adverse Drug Reactions to the Perturbation of Underlying Biological Pathways. PLoS One. 2010; 5:e12063. [PubMed: 20808786]
34. H H. Relations between Two Sets of Variates. Biometrika. 1936; 28:57.
35. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical Correlation Analysis: An Overview with Application to Learning Methods. Neural Comput. 2004; 16:2639–2664. [PubMed: 15516276]
36. Witten DM, Tibshirani R, Hastie T. A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis. Biostatistics. 2009; 10:515–534. [PubMed: 19377034]
37. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A Side Effect Resource to Capture Phenotypic Effects of Drugs. Mol Syst Biol. 2010; 6:343. [PubMed: 20087340]
38. Hughes MR, Malloy PJ, Kieback DG, Kesterson RA, Pike JW, Feldman D, O'Malley BW. Point Mutations in the Human Vitamin D Receptor Gene Associated with Hypocalcemic Rickets. Science. 1988; 242:1702–1705. [PubMed: 2849209]
39. Sobol RW. DNA Polymerase Beta Null Mouse Embryonic Fibroblasts Harbor a Homozygous Null Mutation in DNA Polymerase Iota. DNA Repair (Amst). 2007; 6:3–7. [PubMed: 16979388]
40. Lenhard JM, Furfine ES, Jain RG, Ittoop O, Orband-Miller LA, Blanchard SG, Paulik MA, Weiel JE. Hiv Protease Inhibitors Block Adipogenesis and Increase Lipolysis in Vitro. Antiviral Res. 2000; 47:121–129. [PubMed: 10996400]
41. Azzaoui K, Hamon J, Faller B, Whitebread S, Jacoby E, Bender A, Jenkins JL, Urban L. Modeling Promiscuity Based on in Vitro Safety Pharmacology Profiling Data. ChemMedChem. 2007; 2:874–880. [PubMed: 17492703]
42. Overington JP, Al-Lazikani B, Hopkins AL. How Many Drug Targets Are There? Nat Rev Drug Discov. 2006; 5:993–996. [PubMed: 17139284]
43. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS. Drugbank 4.0: Shedding New Light on Drug Metabolism. Nucleic Acids Res. 2014; 42:D1091–D1097. [PubMed: 24203711]
44. Liu T, Altman RB. Identifying Druggable Targets by Protein Microenvironments Matching: Application to Transcription Factors. CPT Pharmacometrics Syst Pharmacol. 2014; 3:e93. [PubMed: 24452614]
45. Wei L, Altman RB, Chang JT. Using the Radial Distributions of Physical Features to Compare Amino Acid Environments and Align Amino Acid Sequences. Pac Symp Biocomput. 1997:465–476. [PubMed: 9390315]

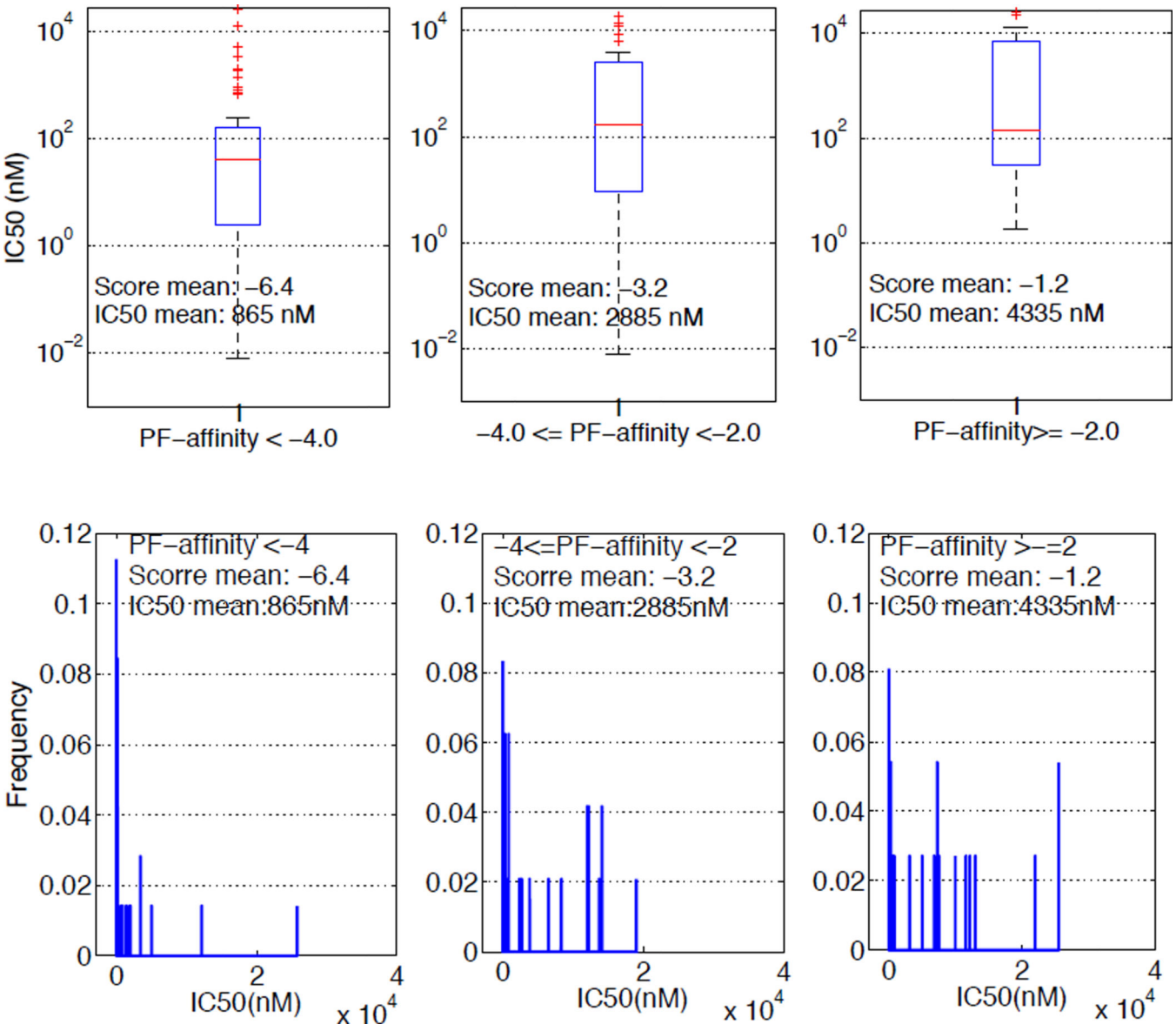
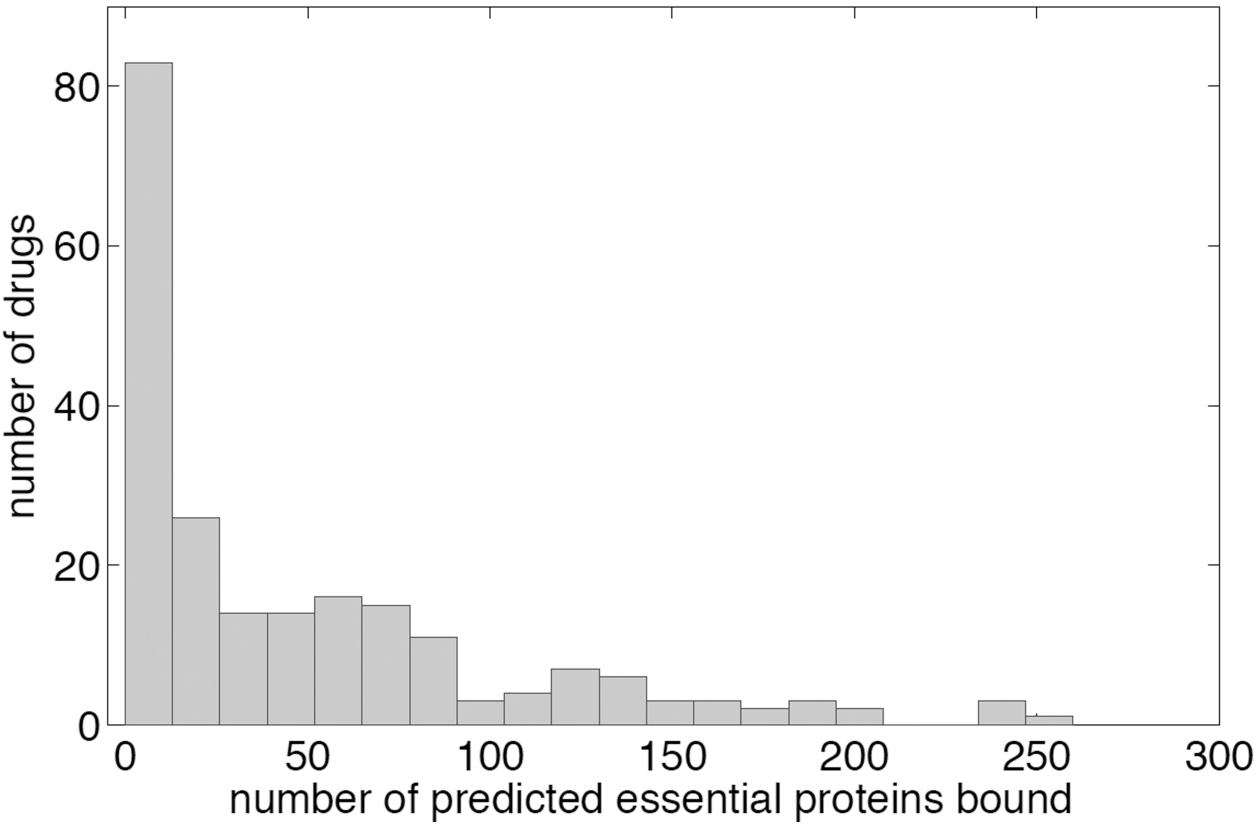


Figure 1.

Predicted PocketFEATURE affinity scores and experimental IC50 values

We calculated the predicted PocketFEATURE affinity scores (PF-affinity) between 216 drugs and 563 proteins, resulting in 121,608 scores between drugs and proteins. We found reliable experimental data from IC50 Assay Dataset for a total of 234 pairs of drug and protein. These 234 pairs are divided into three groups according to the predicted PF-affinities. The corresponding IC50 are shown in boxplots (top panel). On each box, the central mark (red line) is the median (at 95% confidence interval), the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. The bottom panel shows the histogram of IC50 values in different score groups. This figure suggests that the predicted PF-affinity can be used to estimate the potential of binding.

A.



B.

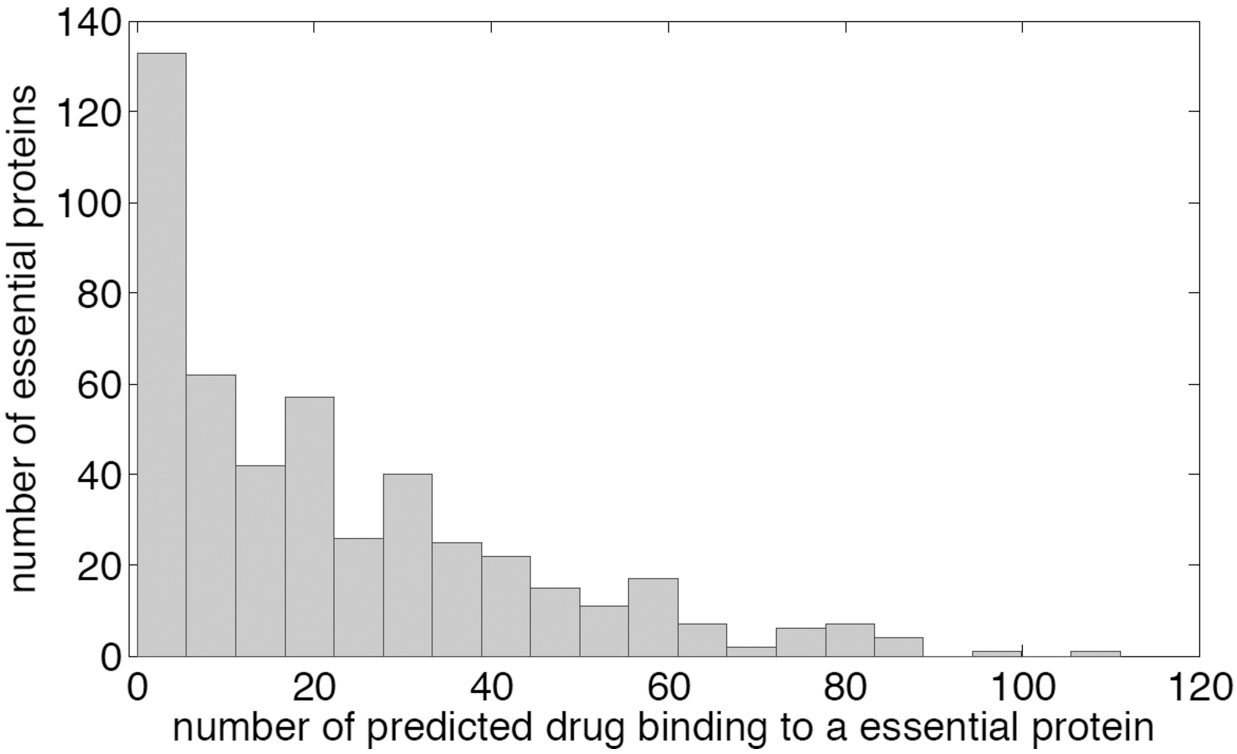


Figure 2.

A. Histogram of the predicted drug promiscuity

For each of the 216 drugs, we estimated its potential for binding each of 563 proteins by calculating the PF-affinity scores. Using an affinity score cutoff of −2.0, we counted the number of predicted essential proteins bound of a given drug. About 15% drugs (34 of 216) bind to 20% or more of the 563 human essential proteins. More than 30% of drugs bind to more than 10% of the essential proteins.

B. Histogram of the predicted target promiscuity

For each of the 563 proteins, we estimated its potential for binding each of the 216 drugs by calculating the PF-affinity scores. Using a score cutoff of −2.0, we counted the number of drugs of a given protein. The most promiscuous proteins are: B-cell CLL/lymphoma 2 (BCL2), E1A binding protein p300(EP300), Janus kinase 2(JAK2), adenylate kinase 3-like 2 (AK4), coagulation factor II (thrombin) receptor (F2R), nuclear receptor subfamily 5 (NR5A2).

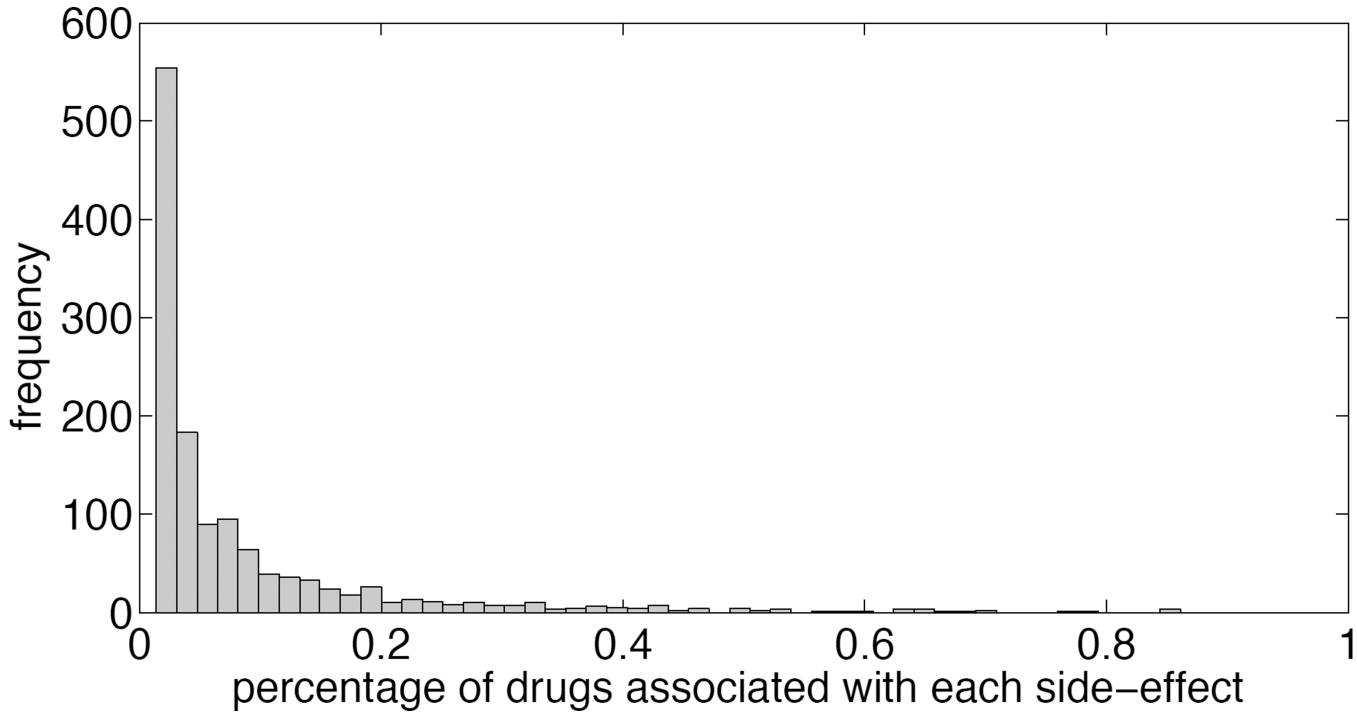


Figure 3.

Histogram of percentage of drugs associated with each of the 1300 side-effects

We have constructed the Drug Binding Dataset that collects 216 small molecule drugs for which 3D conformations of their binding sites are available in PDB and their side-effect are available in SIDER2. After removing side-effects that are observed in two or fewer drugs, the dataset spans 216 drugs and 1300 side-effects.

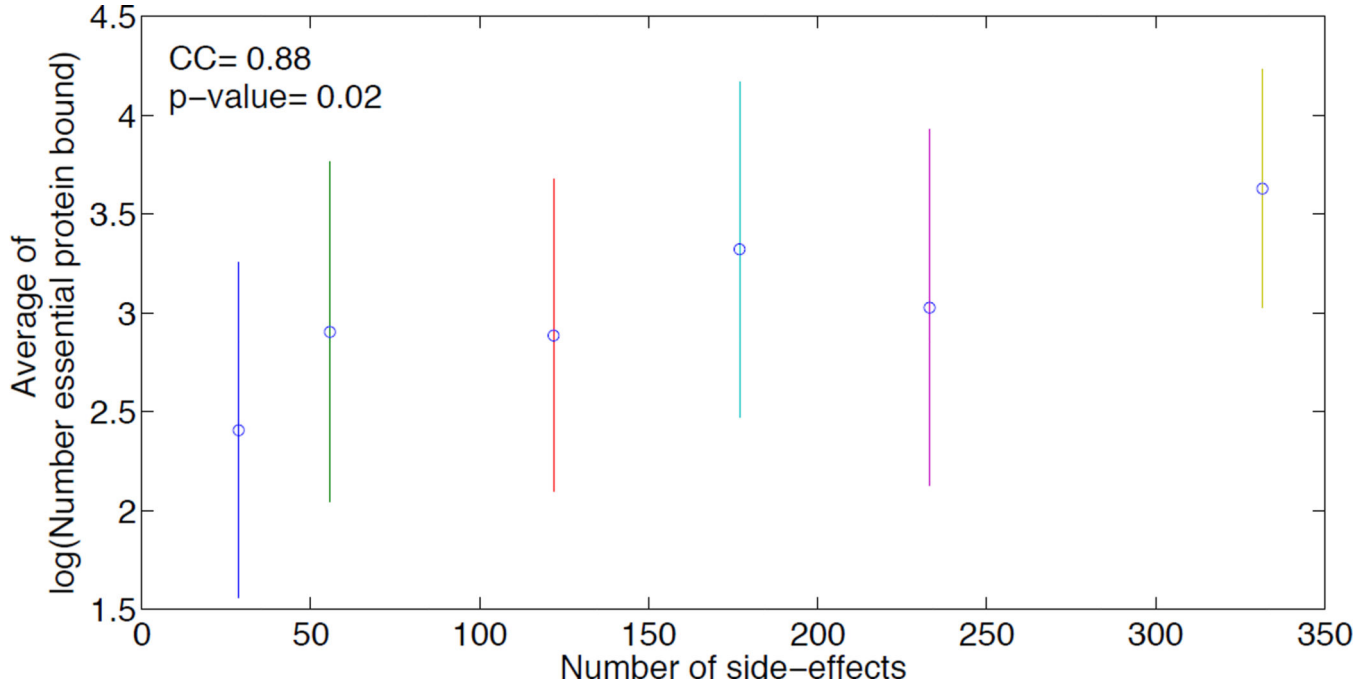


Figure 4.

Relationship between the predicted drug promiscuity and side-effect

For each of the 216 drugs, we estimated its potential for binding each of 563 proteins by calculating the PF-affinity scores. Using an affinity score cutoff of −2.0, we counted the number of predicted essential proteins bound of a given drug. We grouped the 216 drugs according to the number of their observed side-effects. The average number of side-effects in each group correlates with the average number of predicted essential proteins bound.

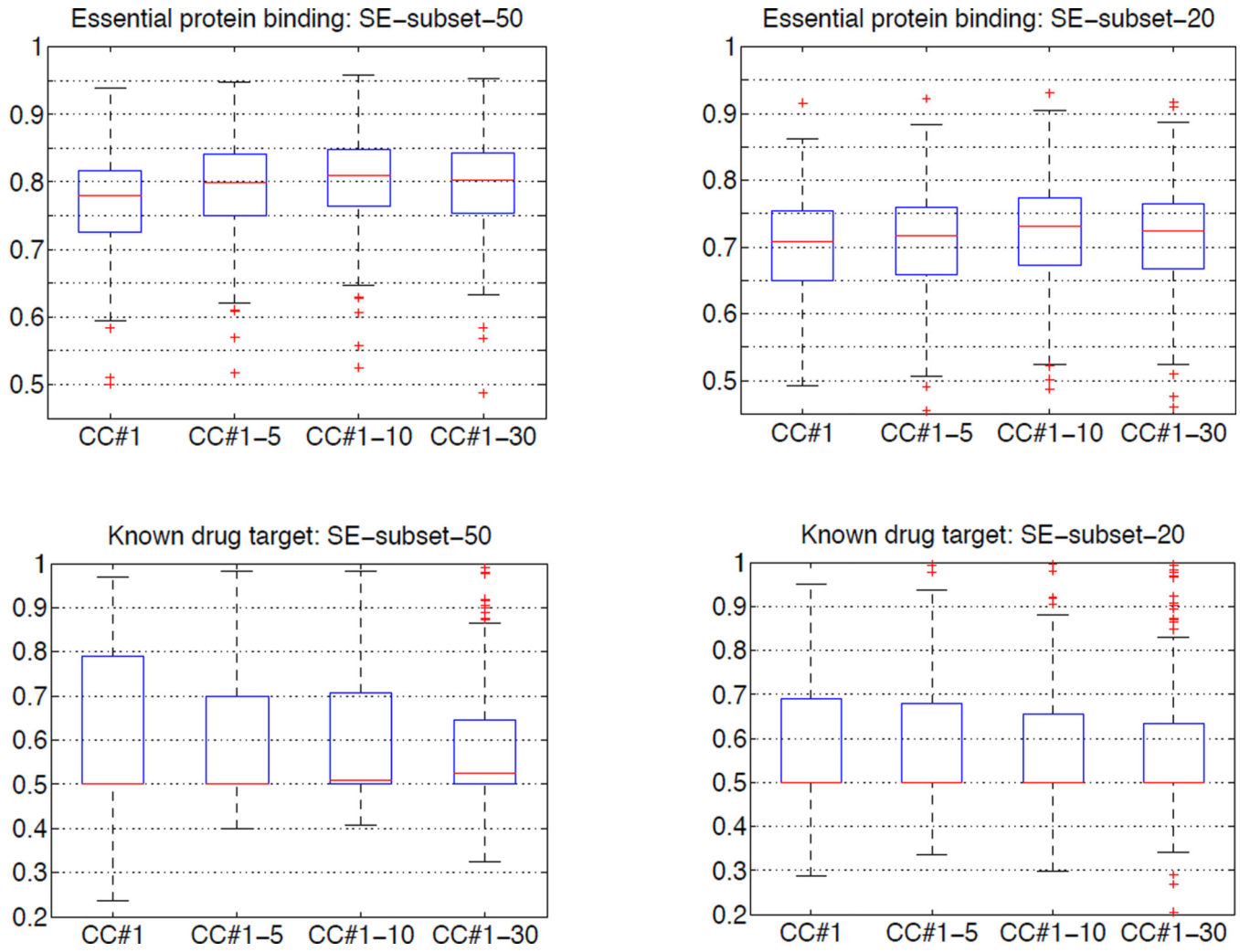
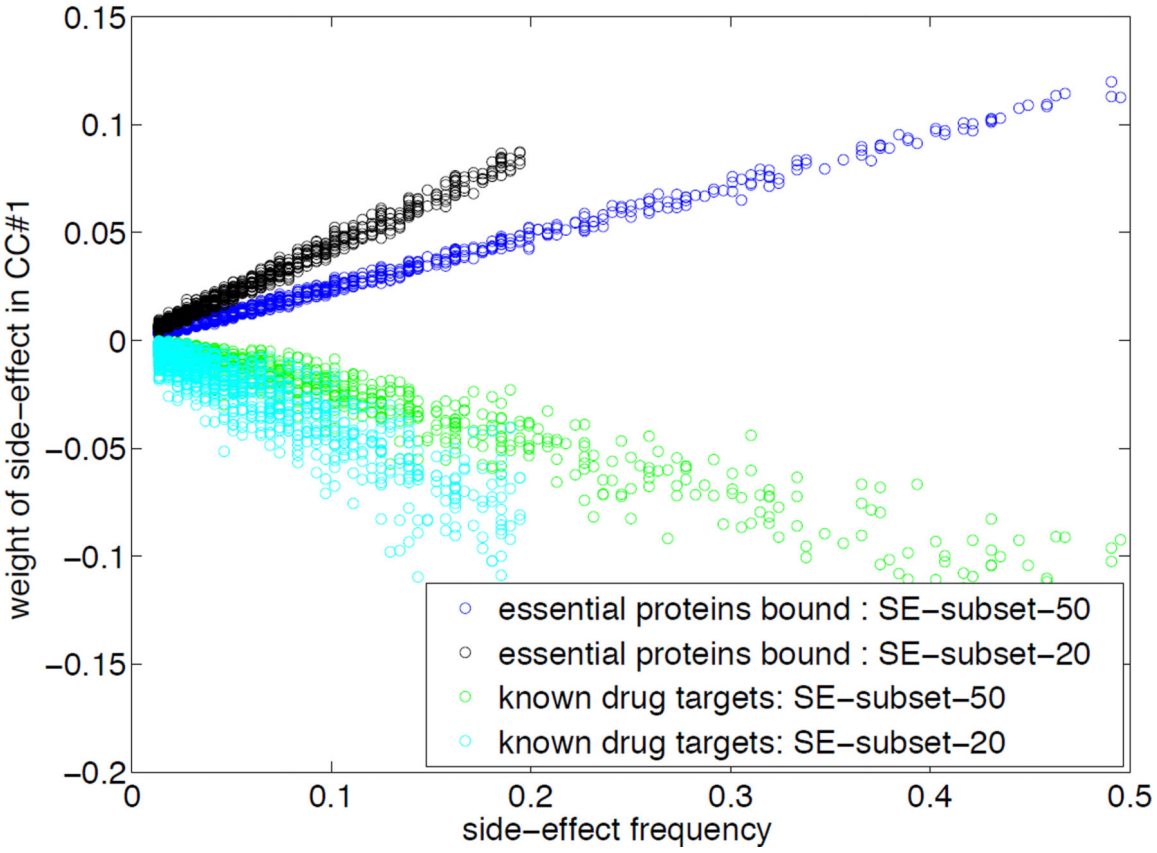


Figure 5.

Performance of SVD-CCA

We have created two subsets of the side-effect data from SIDER: SE-subset-50 includes side-effects that are only observed in fewer than 108 (50%) of the 216 drugs. SE-subset-20 includes side-effects that are only observed in fewer than 43 (20%) of the 216 drugs. We used two different sets of molecular descriptors for drugs: (1) the predicted essential proteins bound (top panel), and (2) the recognized drug targets (bottom panel) in order to compare their performance when predicting side-effects. We evaluated performance by using the AUC scores for side-effect predictions in a leave-one-out cross validation. SVD-CCA has better performance on SE-subset-50, compared to SE-subset-20. Using predicted essential proteins bound leads to higher performance. We further compared the performance of SVD-CCA when using different sets of CCs. When the molecular drug descriptors are predicted essential proteins bound, the average performance is improved as we use more CCs. The best performance is achieved when using the first ten sets of CC (CC#1–10), with the best median AUC of 0.82 for SE-subset-50 and 0.73 for SE-subset-20. However, when using recognized drug targets as the molecular drug descriptor, the performance is best with the first CC and does not improve with additional CCs.

A.



B.

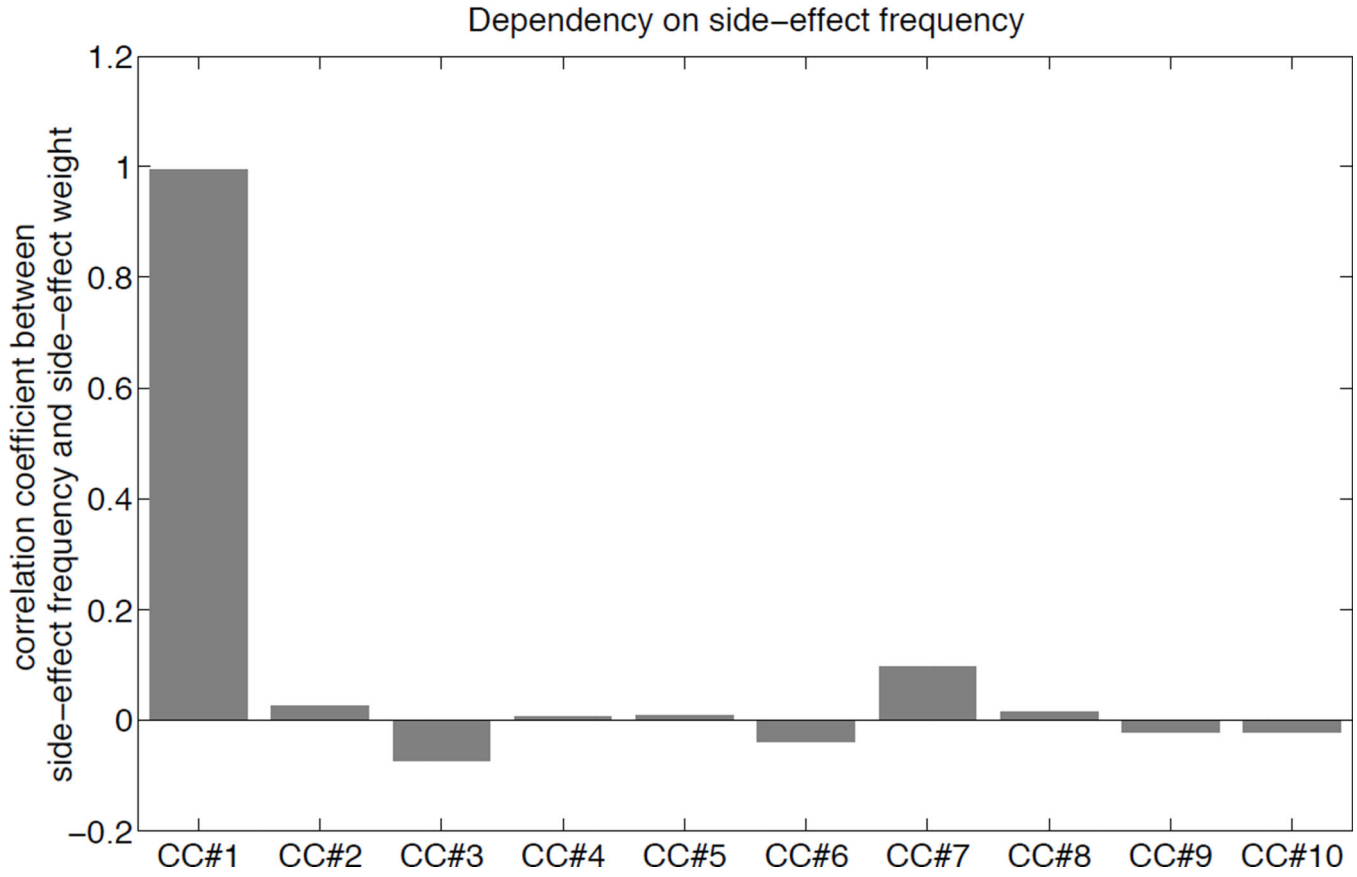


Figure 6.

A. Characters of weight of side-effect assigned in CC#1

We calculated side-effect frequency (the number of drugs associated with each side effect) observed in SE-subset-50 and SE-subset-20. The weights assigned to each of side-effect in the first canonical components are plotted against side-effect frequency. When using essential proteins bound in SVD-CCA, the weight of side-effect in CC#1 correlates with side-effect frequency very well. When using known drug targets in SVD-CCA, the weight of side-effect in CC#1 also shows a reasonable correlation with side-effect frequency.

B. Characters of weight of side-effect assigned in CC#1–10

We calculated correlation coefficient between side-effect frequency observed in SE-subset-20 and the weights assigned to side-effect in each of the first ten canonical components. The weight assigned to side-effect in CC#1 correlates with the side-effect frequency (correlation coefficient 0.99). There is no clear correlation between side-effect frequency and the weight assigned to side-effect in CC#2 to CC#10.

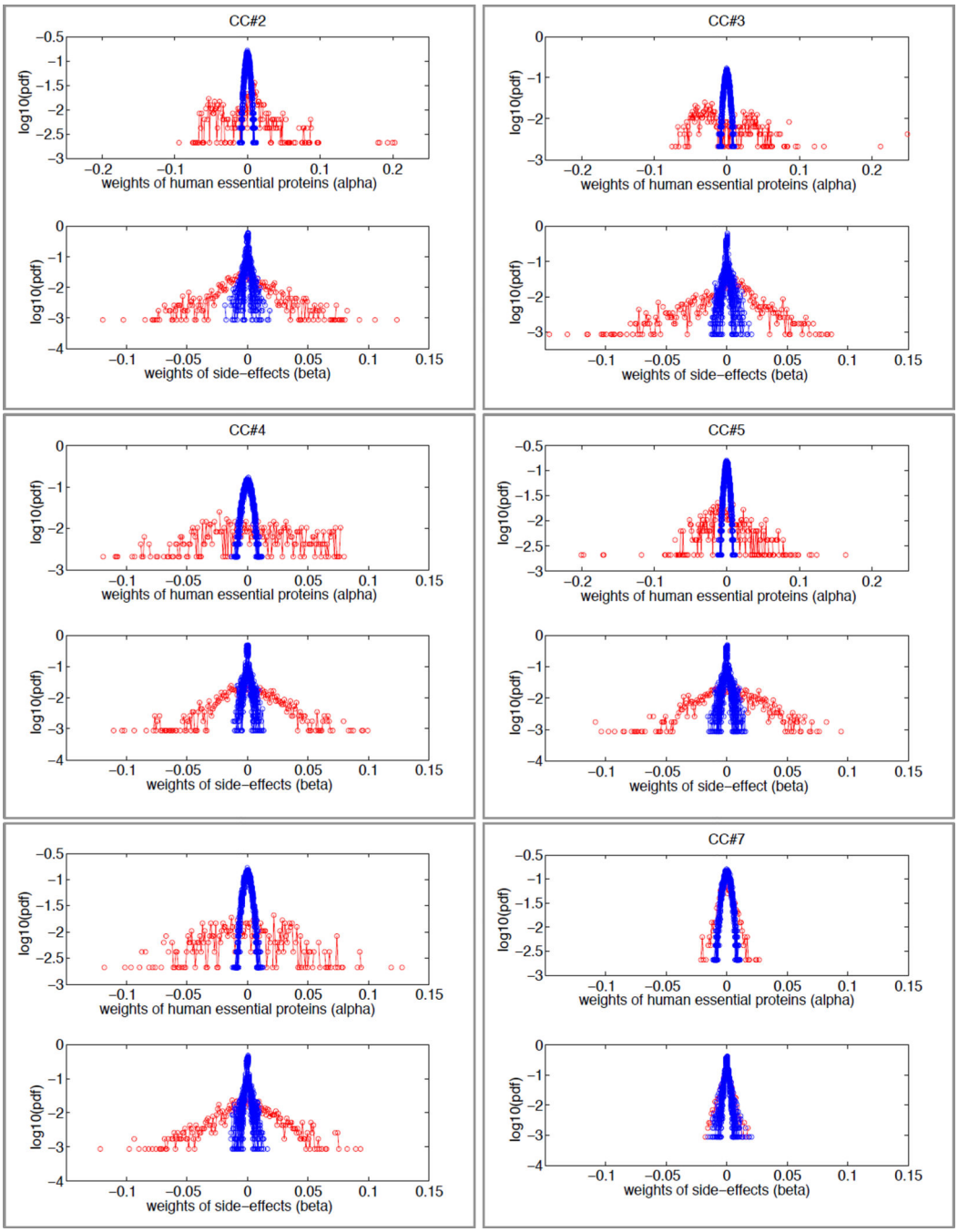


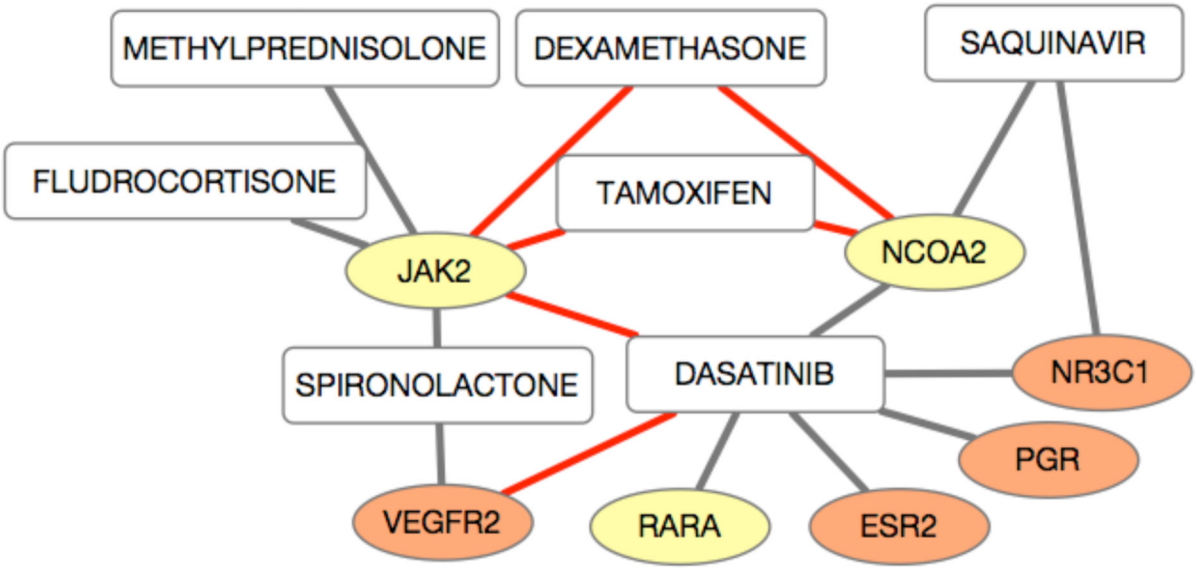
Figure 7.

Significantly weighted attributes

For each canonical component, we extracted weight of each essential protein and side-effect (red circles). We then compared the values to those extracted from the permutation tests of 100 iterations. The probability distributions (pdf) of weight values are shown below. The blue clusters represent the weights derived from 100 permutation tests. In SVD-CCA process, extreme values are assigned to essential proteins and side-effects. The plot suggests

that the extreme values observed in CC#2, CC#3, CC#4, CC#5 and CC#6 are significant, but not in CC#7.

A.



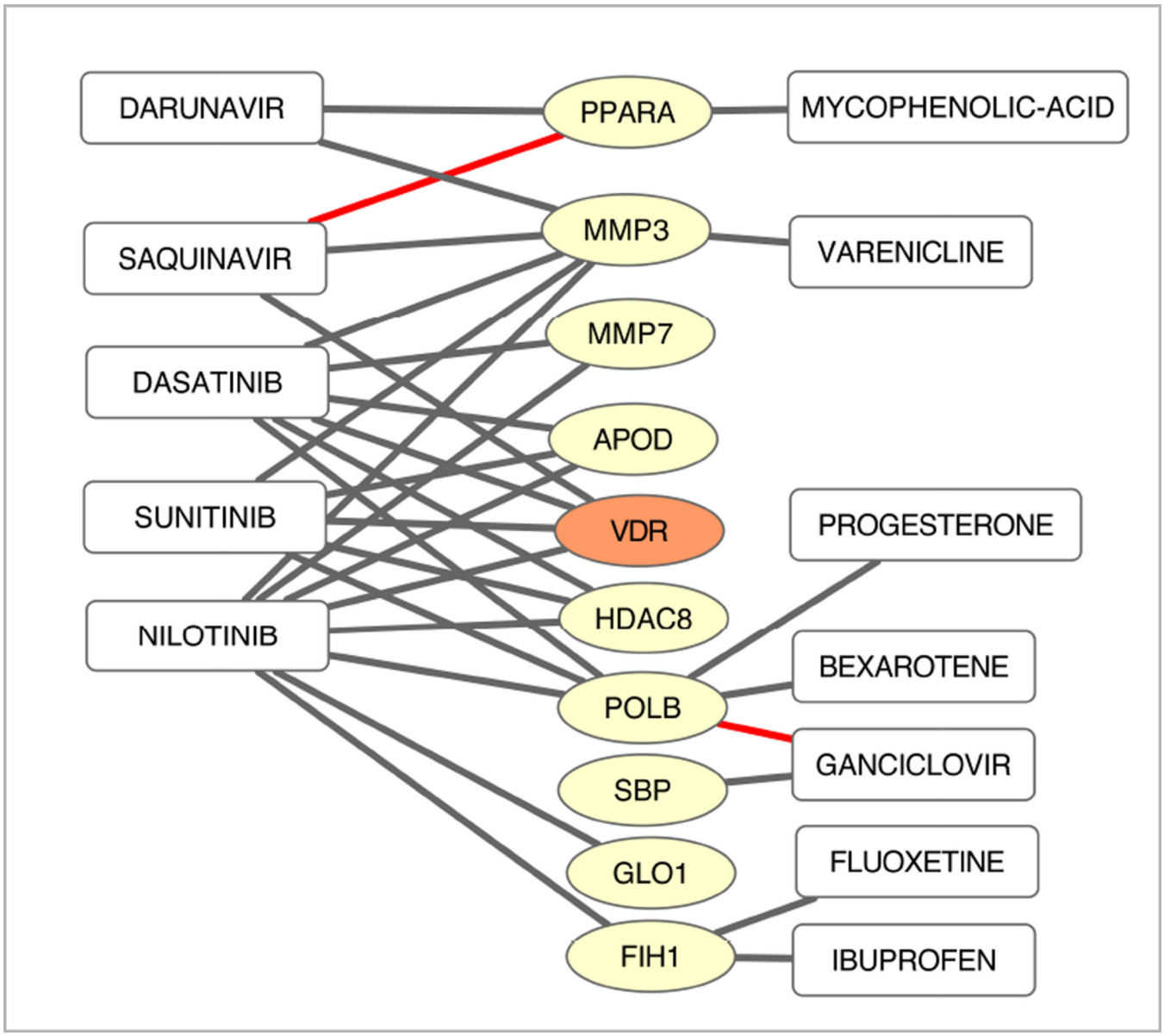


Figure 8.

A. Seven essential proteins are associated with menstrual irregularities/disorders in SVD-CCA analysis. They are: glucocorticoid receptor (NR3C1); glucocorticoid Nuclear Receptor 2(NCOA2), progesterone-receptor (PGR), retinoic-acid-receptor (RARA), estrogen-receptor (ESR2), Janus kinase 2 (JAK2), vascular endothelial growth factor 2 (VEGFR2). Experimental assays in our IC50 Assay Dataset confirmed binding between six pairs of essential proteins and drugs (red edge): tamoxifen to NCOA2 and JAK2, dexamethasone to NCOA2 and JAK2 and dasatinib to JAK2 and VEGFR2. Experiments conducted by Novartis reveal ten proteins associated with menstrual irregularities. Four of them are the same with our predicted ones (orange highlight): estrogen-receptor (ESR2), progesterone-receptor (PGR), glucocorticoid receptor (NR3C1) and vascular endothelial growth factor 2 (VEGFR2).

B. Proteins associated with hypocalcemia in SVD-CCA analysis

They are: apolipoprotein (APOD), DNA polymerase beta (POLB), histone deacetylase 8 (HDAC8), hypoxia-inducible factor 1-alpha inhibitor (FIH1), glyoxalase I (GLO1), matrix metalloproteinase-3 (MMP3), matrix metalloproteinase-7 (MMP7), peroxisome proliferator-activated receptor alpha (PPARA), sex steroid-binding protein (SBP) and vitamin D3 receptor (VDR). One survey reported two proteins to be associated with hypocalcemia: calcium-sensing receptor (CASR) and vitamin D3 receptor(VDR). Naturally occurring mutations in the calcium-sensing receptor gene (CASR) cause hypocalcaemia or hypercalcemia. In knockout mice, genetic inactivation of VDR leads to hypocalcemia. We have predicted VDR as an important protein that may interact with saquinavir (HIV drug) and three cancer drugs. In addition, ganciclovir is known to interact with POLB and saquinavir binds to PPARA. Since hypocalcemia is a side effect for many drugs, the predicted essential proteins bound may provide insights into diverse mechanisms of hypocalcemia.

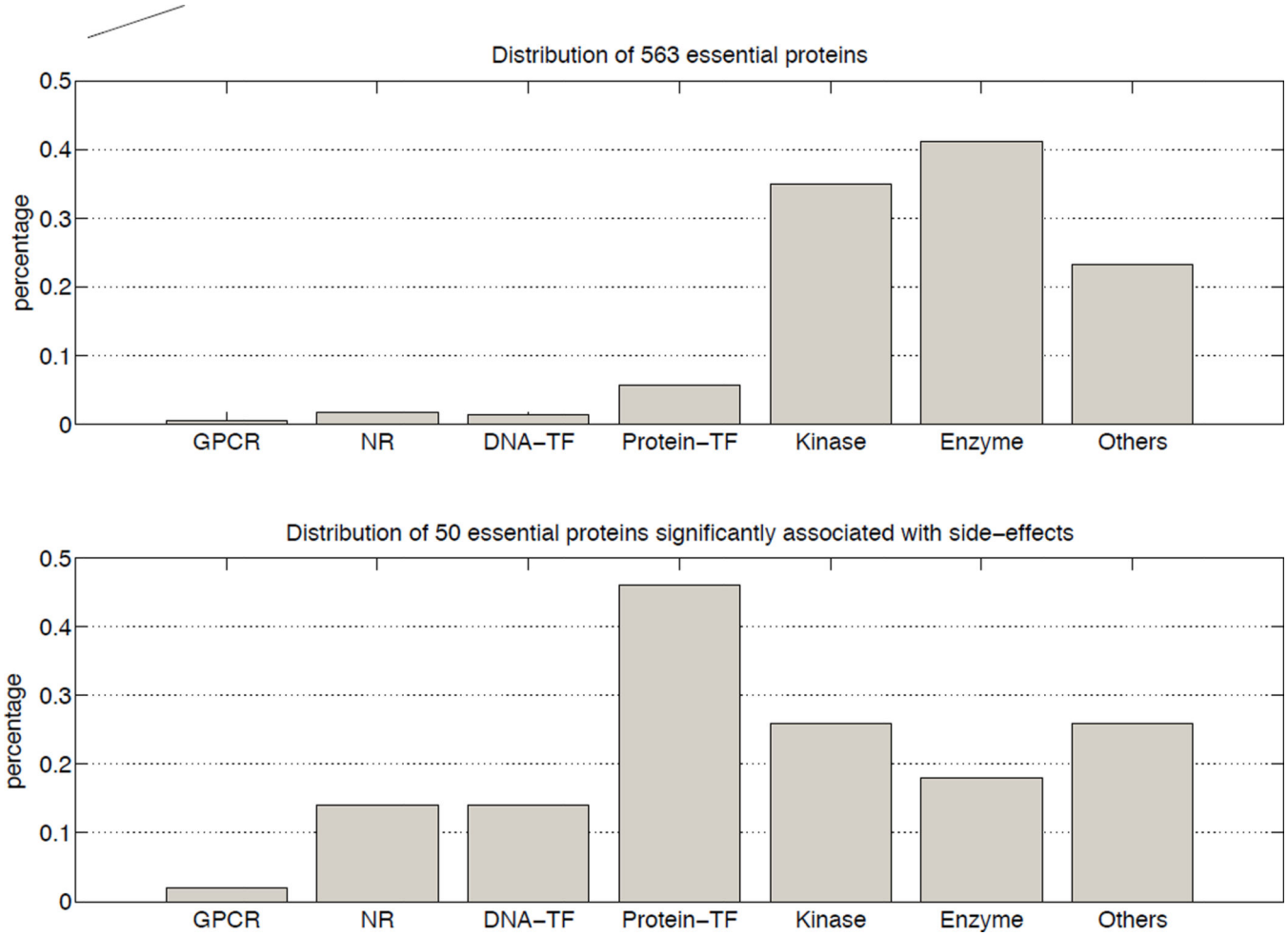


Figure 9.

Molecular functions of essential proteins

Using the Gene Ontology (GO), we grouped the 563 human essential proteins into GPCR (GO:0004930), nuclear receptors (NR)(GO:0004879), DNA-binding transcription factors (GO:0003700), protein-binding transcription factors (GO:0000988), kinases (GO:0016301), or enzymes (GO:003824, excluding kinases) and others. The distributions of the 563 human essential proteins are shown in the top panel and that of the essential proteins significantly associated with drug side-effects are shown in the bottom panel.