Exploring Polypharmacology in Drug Discovery and Repurposing Using the CANDO Platform

Gaurav Chopra1,2,* and Ram Samudrala3,*

1Department of Chemistry, Purdue University, West Lafayette, IN, USA; 2Purdue Center for Drug Discovery, Purdue University, West Lafayette, IN, USA; 3Department of Biomedical Informatics, SUNY, Buffalo, NY, USA

Abstract: Background: Traditional drug discovery approaches focus on a limited set of target molecules for treatment against specific indications/diseases. However, drug absorption, dispersion, metabolism, and excretion (ADME) involve interactions with multiple protein systems. Drugs approved for particular indication(s) may be repurposed as novel therapeutics for others. The severely declining rate of discovery and increasing costs of new drugs illustrate the limitations of the traditional reductionist paradigm in drug discovery. Methods: We developed the Computational Analysis of Novel Drug Opportunities (CANDO) platform based on a hypothesis that drugs function by interacting with multiple protein targets to create a molecular interaction signature that can be exploited for therapeutic repurposing and discovery. We compiled a library of compounds that are human ingestible with minimal side effects, followed by an ‘all-compounds’ vs ‘all-proteins’ fragment-based multitarget docking with dynamics screen to construct compound-proteome interaction matrices that were then analyzed to determine similarity of drug behavior. The proteomic signature similarity of drugs is then ranked to make putative drug predictions for all indications in a shotgun manner. Results: We have previously applied this platform with success in both retrospective benchmarking and prospective validation, and to understand the effect of druggable protein classes on repurposing accuracy. Here we use the CANDO platform to analyze and determine the contribution of multidrugging (polypharmacology) to drug repurposing benchmarking accuracy. Taken together with our previous work, our results indicate that a large number of protein structures with diverse fold space and a specific polypharmacological interactome is necessary for accurate drug predictions using our proteomic and evolutionary drug discovery and repurposing platform. Conclusion: These results have implications for future drug development and repurposing in the context of polypharmacology.

Keywords: Shotgun drug discovery, virtual screening, computational proteomics, polypharmacology, multitarget.

INTRODUCTION

Traditional approaches to drug discovery focus on a limited set of interactions between individual protein targets and small molecule compounds. The goal generally is to target an essential protein responsible for pathogenesis so as to completely inhibit its function. Almost all current drugs have been developed by this approach. However, the number of novel drugs being discovered every year has been reduced to a handful. Currently less than 30 new drugs approved each year and most of them are analogues to other existing drugs or other patent workarounds (data obtained from <fda.gov>) and the estimated average total costs for developing a novel drug and bringing it to market can be up to $2.6 billion, according to the Tufts Center for the Study of Drug Development (<csdd.tufts.edu>). Thus there exists severe limitations for novel drug development as it is typically associated with including long time-frames (10-15 years) and large investment outlays [1-4].

A solution is to repurpose/reposition existing drugs that are relatively benign in terms of side effects for new indications [3, 5-11]. We were one of the first groups to propose shotgun drug repurposing for malaria, using a computational multitarget docking with dynamics approach [9]; we have since validated our predictive models numerous times (see [3, 5-15] for a few examples). Drug repurposing can be made much more efficient and efficacious by considering variations (mutations) in proteins encoded by individual host genomes to identify several candidates for treatment of the same indication/disease paving the path towards exome-specific clinical trial groups, thereby enabling personalized/precision medicine. Systematic exploration of drug repurposing opportunities by sharing preclinical and clinical trials data is hindered by extensive competition within the pharmaceutical industry related to intellectual property issues. We utilize this repurposing paradigm along with a computational platform we have developed that evaluates relationships between compound-proteome interaction signatures to predict genome- and indication-specific drug regimens for particular individuals in a shotgun and holistic manner (i.e., against all indications simultaneously). To verify and improve the accuracy of our platform, we have been working with collaborators worldwide for preclinical and clinical validation. The experimental results obtained are integrated back into the modeling platform to iteratively improve putative drug prediction accuracy. Here, we use this platform to explore drug polypharmacology.

Overview of Drug Discovery

Small molecule drug discovery begins by assaying compounds for in vitro activity against diseases/indications of interest in a high-throughput manner with or without knowledge of molecular targets identified by biochemistry and molecular biology [10, 16, 17]. Compounds observed to be active in a phenotypic (protein or cell) assay (hits) are further tested in vivo, typically in animal models when available, while pharmacokinetic and pharmacodynamic properties are simultaneously measured. Compounds that are successful at this stage are considered leads and go through further biochemical and structural optimization. Once toxicity and efficacy against the disease of interest is established, the candidate molecule is considered for testing in human patients and formally designated as a drug candidate. The candidate drug is then subjected to preclinical and clinical studies [18] and if successful, undergoes commercial production and distribution [19].

Druggable Protein Classes

Many drugs have no known mechanism of action. Even when they are efficacious, the reasons for side effects are not well under-
stood and there is no drug without any side effects whatsoever at the maximal clinically approved doses in all patients. Drug discovery focuses on the elucidation of mechanistic interactions between small molecules and particular “druggable” classes of proteins. G-protein-coupled receptors (GPCRs) and protein kinases are among the most popular classes of targets for a wide range of indications/diseases [20-25].

GPCRs constitute a large superfamily of transmembrane protein receptors unique to eukaryotes and are the most successful targets in modern medicine. Approximately 36% of marketed pharmaceuticals target GPCRs [20]. However, the endogenous ligands of 100-200 GPCRs remain unidentified. These so-called orphan GPCRs with unidentified natural functions are an important source of drug targets [21].

Protein kinases constitute nearly 2% of all human genes, and play a critical role in cellular signaling pathways [22]. Perturbations to these pathways have been linked to cancer, diabetes, and other inflammatory diseases [23]. These perturbations are often due to mutation, translocation, or upregulation events that cause one or more kinases to become highly active and not respond normally to regulatory signals [24]. As a result, much of the effort in developing treatments for these diseases has focused on shutting down these aberrant kinases with targeted inhibitors [25].

Tyrosine kinase inhibitors (TKIs) in particular have shown promise as powerful therapeutics in the treatment of human cancers. Arguably, the biggest success story has been the discovery of imatinib for the treatment of chronic myeloid leukemia (CML), a white blood cell cancer. CML results from a fusion event that produces an active form of the ABL kinase in 95% of CML patients [26]. Imatinib has been highly effective with minimal toxicity in treating patients with CML. It is thought that these properties of imatinib are due to its high degree of selectivity for ABL; imatinib does not bind to the closely related SRC kinase, even though residues at the ABL-imatinib binding interface are nearly identical [26, 27]. Unfortunately, even when perceived highly selective TKIs like imatinib are available, the emergence of resistance mutations limits the duration of therapeutic benefit.

While the mechanisms of some drugs are somewhat well elucidated as illustrated above, our work indicates that most human ingestible drugs function not only by interacting with multiple proteins but also that these proteins are from different druggable protein classes [15]. Current data suggest that protein functions are distributed unevenly across protein structures [28]. The most broadly populated fold families most often catalyze similar chemical functions, but to different substrates. The underlying structural differences can be subtle as in the case of conserved binding sites that bind ATP in protein kinases. Up to 4% of the proteome of higher organisms such as mammals can be represented by a single fold family such as G-protein like receptors, each of which serves as a receptor to recognize a different distribution of ligands. This pattern of different molecular interactions being carried out by proteins with the same fold makes inferring mechanistic difficult. Advances in methodological accuracy and efficiency that enable mapping of the structural differences underlying these different functions are now detectable through iterative superposition of the corresponding domain and binding sites [29-31]. Differences between advancements in knowledge-based and ab initio protein structure prediction may be instructive for advancement by bioinformatic mapping and docking [32]. With an explosion of databases of biochemical and structural interactome data, the field appears ripe for system-wide bioinformatic mapping of compound-proteome interactions.

Drug Resistance and Multitargeting

We and others have previously explored the effect of amino acid mutations in protein structures in the context of drug interactions for infectious indications such as HIV and malaria [33-38]. With respect to neoplastic indications, drug resistance is thought to be the reason for treatment failure in over 90% of patients with metastatic cancer [39]. For CML, the observation of a large number of secondary resistance cases (where patients regressed after several months of effective treatment) prioritized the development of a second generation of specific kinase inhibitors [26, 40]. While there can be many factors for this resistance, 90% of CML patients who relapse after initially positive response to imatinib have mutations in the BCR-ABL kinase domain that disrupt imatinib binding [26, 40-42]. Laboratory and clinical studies suggested that these mutations are not random. Particular mutations were observed with high frequency, indicating a simple physical mechanism driving the emergence of resistance [26, 41, 42]. Furthermore, upon administration of second-line BCR-ABL kinase inhibitors, new resistance mutations against these drugs was observed clinically [43]. All these observations suggest a role for underlying multiple pathways acting in a disease specific manner in response to a synthetic agent, like imatinib, to cause specific mutations for drug resistance. This so-called synthetic lethality may be overcome by designing drugs to work in a disease specific polypharmacological manner by taking into account an interaction a network of multiple disease pathways and drug interactions (Fig. 1).

Imatinib was discovered by high-throughput screening followed by lead optimization [26], a process that is time-consuming, expensive, and largely ineffective due to high failure rates [44]. Imatinib is also non-specific with imperfect selectivity, in that it also inhibits a number of other kinases [45, 46], and is found to be active against constitutively activated tyrosine kinases such as platelet-derived growth factor receptor (PDGF-R), c-Kit, and macrophage colony stimulating factor receptor (c-fms) [47-50]. The mechanism of action of imatinib appears to be multi-faceted and likely involves the inhibition of multiple kinase targets that regulate various pathogenic functions in distinct tissue and cell compartments (Fig. 1). Imatinib treatment can also impair T-cell function in CML patients [51-53]. Rapid design of inhibitors with desired selectivity profiles, potentially targeting multiple kinases simultaneously to combat resistance [54] will require computational approaches that consider the affinity to the entire kinome simultaneously [55, 56].

The multitarget approach is a necessary one because every drug has to be effective at its site of action and also has to be readily metabolized by the body (for instance, by the cytochrome P450 (CYP450) enzymes, which are responsible for metabolizing the majority of drugs). A majority of small molecule drugs are derived from plant sources [18, 57]. Since these molecules are a result of a dynamic interplay of evolution between plants and other organisms sharing their environment, we hypothesize that interesting or functional small molecules that become drugs have multiple modes of action. Computational screening for multitarget binding and inhibition is effective because it exploits the evolutionary fact that protein structure is more conserved than sequence and function, and provides logical evidence that one compound can be an excellent initial candidate for many different protein targets. We have thus developed a platform that is agnostic to how individual compound-protein interactions are determined (whether predicted or observed) but rather relies on an interaction signature which is either a binary or real value set of numbers that indicates how well a compound interacts with a library of protein structures considered representative of the (current) structural universe.

Enabling Multiple Drug Combinations

In many cases, no single drug is sufficiently effective in the therapeutic range to cure a disease, or even to reduce symptoms or recurrence effectively. Thus multiple drugs can be combined to strengthen the effect. Independent effects due to interaction with multiple targets can decrease therapeutic doses, so that less efficacious and slightly more toxic compounds can be used safely and synergistically to achieve the desired efficacy profile. Similar to the
interactions between a receptor (typically, proteins, DNA, RNA, Docking refers to physical three dimensional (3D) structural
Overview of Docking
to efficiencies in cost and labor [62]. In terms of mechanism, a given drug works by binding to one or more
target proteins of interest, but also binds to other proteins causing off-target (neutral) and anti-target (negative) side effects. The CANDO platform makes
interaction predictions for "all drugs" (currently a library of 3733 human approved compounds) against "all proteins" (currently a library of 48,278 proteins) to
determine its efficacy against whole systems of interest, thereby enabling it to make putative drug predictions for “all” indications (that the library of drugs are currently approved for) simultaneously by performing comparative analyses of drug-proteome interaction signatures. In contrast to traditional drug discovery
approaches focused on single targets, our innovative paradigm inverts the traditional one by first compiling a library of compounds/drugs that are safe to ingest with established side effects, and the shotgun virtual screen against whole proteomes/interactomes enables drug predictions for all indications simultaneously.
concept of synthetic lethality in cancer, pathogens often develop resistance to single drug therapy, but simultaneous occurrence of multiple mutations that are resistance to a drug cocktail are exponentially less prevalent [58, 59]. We use the concept of multitargeting, or polypharmacology, where a single dose interacts with more than one target and acts synergistically in a disease specific manner, to address these issues.
Perhaps the most successful application of intentional multitarget drug administration is the use of inhibitors against HIV reverse transcriptase, protease, and integrase in the fight against HIV/AIDS [60]. Multidosing is used in a trial-and-error manner where errors usually result in patient suffering and mortality. In order to avoid such situations during human clinical testing or clinical practice, novel approaches have emerged to model synergistic effects of polypharmacology. For example, combinatorial effects have been tested in vitro using an automated robotics and informatics pipeline. There is an effort to identify combinations of compounds that display synergistic effects, such as inhibiting cytokine storm, as well as tumor inhibition, along with low toxicity and higher efficacy. Such complexity is possible due to either a single protein being targeted by multiple inhibitors, or, more likely, inhibition of multiple proteins involved in the same physiologic process [61]. Such interactions can be modeled and predicted by computational approaches. To this end, docking protocols and methods to predict structure activity relationships (SAR) in the context of multitargeting have been integrated into the drug development process, leading to efficiencies in cost and labor [62].
Overview of Docking
Docking refers to physical three dimensional (3D) structural interactions between a receptor (typically, proteins, DNA, RNA, etc.) and a ligand (small molecules, proteins, peptides, etc.) [3, 9, 10, 14, 63-75]. Docking methods are evaluated by measuring the correct pose/binding mode (using RMSD or TMScore of the coordinates of the atoms) or by measuring predicted affinities [10, 64, 65, 70, 71, 76]. More than 20 molecular docking software tools are currently in use for pharmaceutical research. Autodock Vina [77], Gold [78], and Glide [79] are a few examples of commonly used docking software. Each is based on a variety of cheminformatics, forcefield, and atomic bond flexibility algorithms in predicting poses, and meaningful direct comparison of efficiencies and accuracies is difficult. All have been used successfully in virtual high throughput screening to predict hits and leads ([3, 9, 10], are instances of protocols developed by us). However, the use of computational methods for drug discovery is still in its infancy [75, 80]. We have learnt from our previous experiences [81] and developed a hierarchical fragment-based docking with dynamics algorithm [82, 83] using a generalized all atom potential [74] that exploits evolutionary and structural information to predict the >180 million compound-protein interactions analyzed in this study.
Leveraging Docking for Drug Discovery: The CANDO Platform
Docking methods are primarily designed to predict conformational accuracy. In the context of drug discovery (and in our personal experience at translational research), predicting the conformation accurately is not as important as predicting the binding affinity or, more importantly, functional inhibition. These factors are generally not well considered by traditional docking methods [62]. Further, evaluating global effects of a single compound (or a cocktail) on the proteome is necessary to translate the utility of docking methods to medicine.
The Computational Analysis of Novel Drug Opportunities (CANDO) platform (Fig. 2) predicts the rough poses of compound-proteome interactions bioinformatically and hierarchically refines them using fragment-based docking with dynamics simulations of all the atoms in the system, which we have demonstrated is necessary for accurate calculation of binding energies [81]. Also, we have demonstrated that all-atom knowledge-based force fields are much more accurate and consistent than using classical models of physics based force fields for both protein structure prediction and docking [84-90]. We have shown that using docking with dynamics in a multitargeting manner leads to improved hit rates for finding inhibitors of pathogens compared to conventional approaches [9, 10].

Our CANDO platform evaluates how all (currently 3,733, as of 20 March 2014) FDA approved and other human ingestible drugs (such as dietary supplements) interact with all (currently 48,278, as of 20 March 2014) protein structures from multiple species as a representative of the protein universe (46,784 of which were used in this study). These compound-proteome interaction signatures were generated using a fragment-based docking with dynamics protocol combined with bio- and chem-informatics approaches. The platform then uses similarity of compound-proteome interaction signatures as indicative of similar functional behavior; nonsimilar signatures (or regions of signatures) are indicative of off- and anti-target [91] interactions, in effect inferring homology of compound/drug behavior at a proteomic level. Similarity can be determined using several metrics, from straightforward summations of identical yes/no interaction predictions and (root mean squared) deviations of interactions scores to sophisticated graph theoretic comparisons that consider the similarity of the underlying (actual or predicted) protein-protein interactions compiled from different publicly available methods and databases [92-95] as well as using methods previously developed by us [96-98]. In an evolutionary context, detecting similarity of interaction signatures enables inferring homology of drug...
behavior, and our proposed metrics analyze compound targeting behavior in the context of biologically relevant pathways [92, 93, 99]. The signatures are used to rank compounds for all indications. The platform provides an optimized and enriched set of predicted compound-protein interactions, a comprehensive list of indications and compounds that may be readily repurposed, as well as mechanistic understanding of drug behavior at an atomic level. We have successfully used in this approach to validate potential clinical leads for ten indications over twelve studies, including infectious, autoimmune, oncological, neurological and genetic diseases [11, 100-102] and to understand the contribution of druggable protein classes to benchmarking accuracy [15]. The goal of this study is to determine the contribution of multitargeting/polypharmacology to the accuracy of the CANDO platform.

**APPROACH**

We first describe how compound-proteome interaction signatures are computed and used to rank compounds in an indication-specific manner. We then describe how we retrospectively benchmark drug discovery (as opposed to docking) and repurposing platforms, which is necessary for the machine learning approach to optimize parameters and to make predictions of putative drugs. We finally end with a description of how the CANDO platform was used to analyze and assess the extent to which multitargeting or polypharmacology plays a role in benchmarking (repurposing) accuracy. Our goal is not only to provide a description of the CANDO platform and its accuracy but also to determine contributions to drug behavior in the context of polypharmacology and how that information can be leveraged to develop better drug discovery and repurposing platforms and eventually lead to making more accurate predictions that can be translated to the clinic.

**CANDO Platform and Pipeline**

The CANDO platform is comprised of a unique computational multitarget fragment-based docking with dynamics protocol to implement a comprehensive and efficient drug discovery pipeline with higher efficiency, lowered cost, and increased success rates, compared to current approaches (Fig. 2). The project is funded by 2010 NIH Director’s Pioneer Award and builds upon our extensive work on drug discovery [3, 9-11, 13, 14, 33-36, 74, 81, 103-105], but now applied at the level of the whole proteome/interactome. We completed the first version (v1) of the CANDO platform on 20 March 2014 that resulted in compound-protein interaction matrices representing all compounds interacting with all multi-protein structures representative of the protein universe. The compound library used for screening consists of 3,733 human ingestible compounds including FDA approved drugs and supplements. The initial 39,553 protein structure library (total 48,278 structures) comprises of 14,595 human proteins (8,841 of these are high-confidence models constructed using a combination of one or more protein structure prediction methods including I-TASSER [106, 107], HHBLIT [108], MODELLER [109], KobaMIN [84, 86] and Protinfo CM [110] and 24,958 nonredundant protein structures from all organisms (other than the pathogens explicitly considered) obtained from the protein data bank (PDB).

Based on in-house benchmarking, we have set up a structural modeling pipeline using HHBLITS, I-TASSER and KobaMIN for protein modeling and use COFACTOR [29] and our in-house docking program for initial modeling of small molecule protein interactions. HHBLITSS/TASSER are used to select templates for modeling. I-TASSER iteratively refines the models and templates and our program KobaMIN refines the final models. Protein structure prediction methods are assessed in a blind fashion every two years at the Critical Assessment of Structure Prediction (CASP) experiments. The methods in our modeling pipeline have been among the top performing methods at past CASP experiments. This pipeline has been applied to all human proteins that can be modeled with high accuracy (~60% of the proteome). The benchmarking of individual components (structure prediction, binding site prediction and analysis, docking, potential functions) used in the CANDO platform has been done extensively and published by us and by others [74, 84-90, 106-132].

The pipeline initially uses a bioinformatic docking approach to predict interactions between all the protein structures and all the small molecule compounds. Additional proteins and compounds are added as needed and the CANDO v1 matrix currently consists of an additional 8,745 structures from 17 pathogens (~20% are high confidence models) resulting in final dimensions of 3,733 compounds × 48,278 protein structures or a total of 180,221,774 predicted interactions between proteins and small molecules (the actual number of equivalent docking calculations is greater by almost a factor of 3, representing the average the number of domains and binding sites per protein).

**Indication-Specific Interaction Signature Analysis and Comparison Approach to Ranking Compounds**

The CANDO pipeline uses similarity of interaction signatures across all proteins as indicative of similar functional behavior and nonsimilar signatures (or regions of signatures) as indicative of off- and anti-target [91] interactions, in effect inferring homology of compound/drug behavior at a proteomic level. The main metric to determine similarity is the root mean square deviation (RMSD) of the pair of real value vectors that comprise an interaction signature. Compound-proteome interaction signatures are compared to each other and ranked using protocols that are disease/indication-specific, i.e., compounds approved for particular indications (if available) are used to generate rankings of other compounds that may have similar interaction signatures. Similar compound-proteome interaction signatures are further clustered and protein (target, anti-target, and off-target) specific weighting is introduced wherever information about involvement of particular proteins whose modulation (typically inhibition) is implicated in pathogenesis.

A consensus of the top-ranked indication-specific compound-proteome interaction signatures are evaluated in the context of additional biological information that includes all-atom fragment-based docking with dynamics simulations using knowledge-based potentials developed by us [74, 84, 85, 87] as well as by others [75], the scientific literature, pathways, gene structure and expression data and other information. The top compounds produced by these integrated rankings are considered to be putative repurposed drugs for a particular disease/indication. The platform thus provides an optimized and enriched set of compound-protein interactions, a complete and comprehensive list of indications and compounds that may be readily repurposable, as well as a mechanistic understanding of drug behavior at three dimensional (3D) atomic and molecular scales.

The top ranking predictions are validated at the bench using a barrage of approaches. Based on our extensive experience in predicting drug hits and leads, some of which have been patented and are being licensed commercially <www.fasq.org/patents/app/2011012031>, obtaining atomic level mechanistic understanding is limited in use and approaches that predict functional inhibition are typically most useful in the context of drug discovery.

**Machine Learning Approaches to Iteratively Refine Predictions**

Our approach is to validate the predictions made and obtain bench results that will then iteratively feedback into our pipeline to improve successive predictions. The goal is to compare the results from bench validation for a given set of proteins to the predictions and re-parameterize our platform using machine learning. This enables us to determine which parameter choices have the greatest individual accuracy and also the choices that will lead to compounds becoming viable drugs. Machine learning approaches such
as neural networks and support vector machines (SVMs) are capable of weighting the contribution of individual proteins and compounds and particular regions of the interaction signatures as needed to optimize predictive value. In past work, we have used both neural network and SVM packages written by others to develop methods to predict protein function ([121, 133, 134], for instance).

The benchmarking is coupled with machine learning to determine the various parameter weightings. Our current accuracies using the “All” protein set (48,278 structures in CANDO v1) range from ~12% to up to ~50% depending on the cutoff (top10 to top100) or the number of indications considered (i.e., all 1439 indications with two or more approved compounds or only those ~700 reporting a nonzero result) using only the compound-centric approach to prediction (i.e., all proteins weighted equally). This suggests there is significant room for improvement particularly in the number of indications applicable to the CANDO platform (i.e., the coverage). The greater the coverage, the higher the number of accurate predictions that can be made for all the 2030 indications associated with set of the 3,733 compounds based on mappings obtained from the Comprehensive Toxicology Database <ctdbase.org>.

**Benchmarking Accuracy of Drug Discovery Platforms**

Benchmarking computational platforms for drug discovery is different from benchmarking the ability of software to correctly dock or bind small molecules to proteins. For the latter, the overall accuracy is based on comparisons of predictions to different types of benchmark validation studies. For the former, however, the goal is to determine the ability of a platform to recognize the likelihood that any given prediction of a small molecule will be effective as a drug for a particular indication.

Initially, a leave-one-out benchmark experiment is performed to calibrate the machine learning approach for all indications with more than one approved compound (1439 such indications for the CANDO v1 library). The indication assignment for one of the compounds is “left out” and the interaction signature similarity comparison approach is used to determine the rank of another compound with the same indication and checking whether the compound without the assigned indication is among the top ranked. Each of the compounds is thus compared and ranked by similarity to all the other compounds in our library. The rank(s) of the compound(s) associated with the same indication is used to determine the performance of our approach. The lower the (average) rank, the better the performance. Similarity is determined using several metrics, such as straightforward summations of identical yes/no interaction predictions and (root mean squared) deviations of interactions scores.

The sensitivity and specificity of this approach is then evaluated using receiver operator characteristic [135] curves. As a computational control, fully randomized CANDO matrices (where all the rows and columns, representing compounds and proteins respectively, are moved to randomly selected locations) are used to perform the same benchmarking analysis to determine the type of results that could be obtained by chance. This allows us to assign a probability value that assesses the likelihood of obtaining any particular prediction relative to a random control.

This benchmark test is made more rigorous by using more sophisticated jack-knifing to leave out related compounds, and different variations of these bootstrapping methods are used to minimize overtraining and other biases inherent in the data sources or methods. In the end, a machine learning model based on neural networks and/or SVMs is obtained so that given a CANDO matrix and a compound-indication mapping table as input, the assignments of compounds to indications by interaction signature comparison is optimized. The initial best machine learning models are further iteratively optimized as prospective predictions are validated and the corresponding bench results are fed back into the machine learning process.

The benchmarking described above allows us to (i) better weight the parameters used for various aspects of the structure modeling pipeline; (ii) determine which docking methods, interaction types, and indications are likely to be most accurate and enable assignment of confidence values to each of our predictions (as well as an overall p value); and (iii) optimize accuracy of the entire platform for ranking putative therapeutics. The judicious application of this approach while rigorously minimizing overtraining enables us to bootstrap a platform initially based on docking to one targeted towards drug discovery. The different validations for particular compound-protein interactions are coordinated to maximize the amount of information available for particular high-confidence subsets.

**Contribution of Multitargeting/Polypharmacology to Benchmarking Accuracy**

To assess the contribution of multitargeting/polypharmacology on CANDO benchmarking accuracy, we compared the distributions and trends of compound-protein interactions for all the 3,733 human ingestible compounds in our library for different interaction score cutoffs and using different proteome sets (Fig. 3). Polypharmacology or multitargeting can be defined by multiple interactions of drugs with all known targets for an indication/disease and/or the interactions with multiple targets of different indications/diseases. Our CANDO methodology was developed to identify similarity of compound behavior using multigorganism compound-proteome interaction signatures as a surrogate for compound perturbation of biological function. Our hypothesis is that the biological functions of compounds/drugs are related to the identification of multitargeting profiles that includes target, off-target and anti-target interactions, and that this information needs to be properly utilized for efficient drug discovery and repurposing. The multigorganism proteome data sets considered are labeled ALL for all 46,784 protein structures used in this study, HUMAN for the 14,595 proteins obtained from a reference Homo sapiens proteome that includes 5754 solved structures from the PDB and 8839 computationally modeled structures (described previously), and NRPDB for a set of 24,959 non-redundant solved PDB structures obtained from eukaryotic, prokaryotic, archaea and viral organismal proteomes. We then assess the relationship between the degree of predicted multitargeting/polypharmacology to the benchmarking accuracies obtained for the indications associated with the human use drugs (Fig. 4). Finally, we provide lists of representative indications and respective benchmarking accuracies corresponding to five compounds with the highest degree of predicted multitargeting and five with the lowest (Fig. 5).

**RESULTS AND DISCUSSION**

**Benchmarking the CANDO Platform**

The individual modeling components (structure prediction, binding site prediction and analysis, docking, potential functions) used in the CANDO platform have been benchmarked extensively by us and by others (see methods). The platform itself has been benchmarked using a compound-centric leave-one-out procedure for all indications with at least one approved compound [11, 15]. Briefly summarizing: There are 749/1439 indications where the predictions identify a related compound with the same indication in the top 1% of the ranks (on average). Consistent predictions are produced regardless of the comparison method, metric, or matrix used. In contrast, when randomized compound-proteome matrices are constructed (by randomly swapping the rows and columns representing the compound-proteome interaction signatures) and used for making predictions, only 10-20 out of 1439 indications “work” (using the same criteria as before), with largely inconsistent sets of predictions.
Using fully randomized matrices as computational controls automatically accounts for normalization issues to account for different number of approved drugs for each indication. Random controls are essential as the leave-one-out benchmarking method is more likely to work by chance on indications with tens of approved compounds compared to those with only a few approved compounds. Our benchmarking results indicate higher average accuracies for indications with larger number of approved drugs. Moreover, compared to random controls, the increase in benchmarking accuracy across all indications is much higher than expected by chance: The average of average accuracies obtained using matrices based on variations of the 48,278 proteins is 15.1% and the maximum average accuracy achieved (by filtering for number of compounds) is 49.8%; in contrast the best performing random matrix (out of a 1000) produced an average accuracy of 0.4% and the maximum accuracy achieved when measured as a function of the number of compounds is 0.9%. Based on the benchmarking, we can determine the indications for which the current version of the CANDO platform is most relevant, which enables us to assign confidence values to each of our predictions.

Multitargeting in Human Approved Drugs Using the CANDO Platform

The benchmarking accuracy is a measure of drug repurposeability. To quantify the level of multitargeting/polypharmacology for each compound, we use multiple cutoff scores (>1.5, >1.6, >1.7, >1.8, >1.9) based on our scoring scheme [15] that correspond to moderate to high compound-protein interaction prediction accuracies. All the 3,733 compounds are sorted by their degree of multitargeting based on the highest cutoff score of >1.9 and plotted against the number of interacting proteins (Fig. 3).

Using the ALL (46,784) proteome set we find that the most predicted multitargeting/polypharmacological drugs interact with 464, 298, 190, 123, and 76 proteins using interaction score cutoffs >1.5, >1.6, >1.7, >1.8, >1.9 respectively (Fig. 3a). If we consider only human proteins (Fig. 3b), the most predicted multitargeting drugs interact with 189, 117, 73, 45, and 26 proteins with interaction scores >1.5, >1.6, >1.7, >1.8 and >1.9 respectively. Similarly, for NRPDB (Fig. 3c), we get 263, 174, 113, 75, and 48 interacting proteins for the most predicted multitargeting drugs. Our results indicate that even if our interaction prediction method is...
wrong half the time, there are still a significant number of drugs approved for human use that are highly polypharmacological (Fig. 3). This suggests that this is an essential feature for a small molecule compound becoming a drug as almost all human approved and ingestible compounds exhibit some degree of polypharmacological signal. Some of the most promiscuous drugs using the interaction profile for the ALL proteome set include eptifibatide, a drug used during heart attack or angioplasty to avoid blood clots [136, 137] and belongs to the same class of drugs as aspirin, and the second most promiscuous drug desmopressin, a synthetic replacement for vasopressin, the hormone that is used to treat the symptoms of diabetes insipidus by reducing urine production, and has been suggested as a therapeutic for various other indications (see Fig. 5a) [138, 139].

Next, we assessed the similarity of distributions of multitargeting/polypharmacology for all 3,733 compounds between different proteomes: ALL, HUMAN and NRPDB. Visually the distributions of multitargeting/polypharmacology looks very similar for the different proteomes (Figs. 3a,b,c). Moreover, on average, all drugs interact with a similar (~2.5-3%) fraction of interacting proteins relative to the proteome size (Fig. 3d). However, a two-sample Kolmogorov-Smirnov (K-S) statistical test using a two-sided alternative hypothesis (which is a statistical measure of assessing similarity or difference between two distributions with lower p-value indicating greater difference) shows that ALL is very different from NRPDB as well as from HUMAN, with p-values of 2.2e-16 for all score cutoffs shown in Fig. 3 with an exception of a slightly higher p-value, but still significantly different distribution between ALL and NRPDB for the >1.5 cutoff. Thus, we conclude that almost all human approved drugs are polypharmacological but the particular distributions of interacting proteins are selective and specific. This could be the result of an evolutionary feature that allows for small molecules with a fairly limited set of atoms (tens to hundreds) to confer selectivity and specificity in the context of whole proteomes, and in the context of simultaneously competing with other molecules for binding and performing their biological function.

Relationship Between Multitargeting and Retrospective Drug Discovery Using the CANDO Platform

From the results described in the previous section and Fig. (3), we concluded that most, if not all, human approved drugs are multitargeting, and offered an evolutionary explanation for how small molecule ligands (drugs or otherwise) derive selectivity and specificity in a proteomic context. Here we wanted to assess the degree
Fig. (5). Relationship between indication type and polypharmacology. Panels (a, b) show the top five most and bottom 5 least multitargeting/polypharmacological drugs predicted by the CANDO platform sorted by decreasing order of predicted number of protein interactions. The indications treated by these compounds are listed and the CANDO top10 percent accuracy (retrieval accuracy of the CANDO platform of known drugs for a particular indication in the top 10 ranked list of predictions obtained using compound-proteome signature similarity) is highlighted using colored tiles. There is no simple correlation between multitargeting/polypharmacology and CANDO accuracy for all 3,733 compounds (Fig. 4) but the most multitargeting drugs treat diseases such as B-cell lymphoma, graft vs host disease, lung, kidney, breast, and prostate cancers along with several infectious indications. In contrast, the least multitargeting/polypharmacological drugs are either vitamin supplements or active forms of Vitamin B group, and treat hypertension, vitamin deficiency, nutrition and anemic disorders, and visually contribute lower to drug retrieval accuracy than most multitargeting drugs. Panel (c) shows the indications treated by imatinib, one of the few well-known polypharmacological drugs. Imatinib treats neoplastic diseases, most of which show higher CANDO drug retrieval accuracy than the least polypharmacological drugs, shown in panels (b, c). The black filled circles on each bar represent the number of drugs known to treat the indications also treated by imatinib. Panel (d) shows the relationship between the number of treated indications by most and least multitargeting drugs from panels (a) and (b) respectively, in comparison to imatinib. The bars are colored by the number of predicted interacting proteins that is also indicated on the top of each bar of the sorted list of multitargeting drugs. There seems to be no relation between the number of indications treated by multitargeting drugs and the extent of multitargeting, defined as the number of predicted protein interactions. Altogether, these results indicate the complex nature of polypharmacology and its contribution to drug repurposing, which extends beyond interaction with multiple proteins: imatinib with much lower predicted number of protein interactions treats neoplastic indications similar to the most multitargeting compounds and exhibits a greater repurposing potential based on the number of indications it is approved for treatment.
to which multitargeting/polypharmacology was responsible for CANDO benchmarking accuracy, i.e., whether the drug repurposing signal in CANDO is dominated by the corresponding polypharmacological signal.

Fig. (4) compares the entire distribution of average CANDO benchmarking accuracies for all 1439 indications as a function of polypharmacology, with the order of compounds sorted by the degree of interactions based on the interaction score cutoff of >1.9 (i.e., same order as in Fig. 3). The CANDO accuracy shown is based on the top10 criterion, i.e. the accuracy of retrospectively identifying a known drug-indication association in a ranked list of the top 10 predictions made by the platform. We used the most stringent criterion (top10) to evaluate drug-indication accuracies in a manner that makes them readily verifiable with straightforward bench experiments, as used by us to prospectively verify predictions for 12 different indications.

Fig. (4a-c) shows that multitargeting/polypharmacology does not correlate with CANDO benchmarking accuracy for different proteome sets considered. For the sorted list of compounds, the average CANDO accuracy over all compounds is lower for HUMAN compared to NRPDB and ALL proteins set (Fig. 4d, horizontal lines). The Q-Q plot (Fig. 4d) shows that the distribution of average CANDO benchmarking accuracies for more multitargeting/polypharmacological drugs is not similar between HUMAN-NRPDB and HUMAN-ALL proteome sets. Furthermore, the sorted distribution of the fraction of the maximum number of interacting proteins for each proteome set correlates with average CANDO benchmarking accuracy (Fig. 4d). This suggests that drug discovery is not a simple function of multitargeting/polypharmacology (Fig. 4a-c) but taken together with the results in Fig. 4d and our prior work [11, 15] indicates that the specific distribution of polypharmacology is important for drug repurposing accuracy.

Our results suggest that multitargeting/polypharmacology is a necessary but not sufficient condition for drug discovery. Polypharmacology of approved drugs is likely a real phenomenon, as detailed in the introduction sections and the predictions observed in Fig. (3). That is, most drugs likely work by interacting with a significant number of proteins (tens or hundreds): examples include multi-kinase inhibitor drugs like imatinib, among others. CANDO exploits this feature in the form of interaction signatures for drug discovery and repurposing. Furthermore, polypharmacology refers to positive interactions (binding, inhibition, etc.) to multiple targets, whereas the CANDO platform considers both interacting and non-interacting proteins to define the signature of a compound as a vector of real value scores that quantify strength of interactions. Our findings thus indicate that for drug efficacy, it is not only the significant numbers of protein interactions that are important even in the context of multitargeting/polypharmacology, but also the much larger number of non-interacting proteins and their distribution.

Hence, multitargeting/polypharmacology, in its simplest form as shown in Fig. (3) that does not consider targets, off-targets, and anti-targets, is not sufficient for effective drug discovery and repurposing. A complete description of the polypharmacological signal appears to be essential for defining when small molecules become drugs, how they derive specificity against an indication/disease, and how they play a selective and specific role in modulating metabolic, regulatory, and signaling pathways.

A simple test of using a binary interaction matrix (interaction on or off) results in 50% lower accuracy compared to using the real value interaction scores for benchmarking the CANDO platform (holding all other conditions the same). Furthermore, the real value interaction scores may account for an evolutionary phenomenon where multiple small molecules compete in terms of binding to a single site to influence molecular function, as well as binding to homologous protein molecules across multi-organism proteomes cooperatively to influence biological function. Taken together, this suggests that quantifying the nature of interaction as a competitive/cooperative polypharmacological signal is a better surrogate for polypharmacology compared to using binary yes/no interactions of compounds with proteins.

**Nature of Polypharmacological Interactions to Identify Drugs Efficiently Using the CANDO Platform**

Small molecule drugs work by considering both positive and negative interactions, specific for the indication of interest; defined as the target, off-target and anti-target interactions in the context of one or more indications (Figs. 1 and 2). Thus a simple model of polypharmacology as multitargeting, as shown in Fig. (3), is a very rough proxy, and it is not sufficient for drug discovery and repurposing, as indicated by the data in Fig. (4). At the simplest level, two drugs binding to the same target do not necessarily work synergistically -- if they bind to the same site (mutually exclusively), they are additive [140] and if they bind to distinct sites independently (i.e., without affecting each other's binding), they should be synergistic. So the effects of two compounds on a given target will depend on where they bind, its binding affinity, and mode of action due to binding (agonist/antagonist). Moreover, if one compound binds to multiple proteins, it depends on where it binds, with what affinity, and what action is performed in the context of the indication/disease. Finally, there are many small molecules in a biological system exerting their influence on both molecular and holistic functional interaction simultaneously as well as sequentially. If polypharmacology in its simplest form is considered only to be the strongest positive interactions to a small number of targets, then its applicability for computational drug repurposing is limited. However, a holistic approach that considers positive, negative, and neutral interactions across whole proteomes is likely to be useful, as these interactions are derived from evolutionary methods that use structural homology to derive interaction scores that can likely be used as a surrogate for binding affinity. The benchmarking results obtained using the CANDO platform indicate the use of real value interaction scores over binary on/off interaction scores.

A specific interaction signature for a compound that specifies binding efficacy, nature of interaction, and the secondary effect of neighboring interactions between interacting proteins themselves, defines a continuum of interactions across multiple protein classes. Such an interaction signature is a surrogate for polypharmacology that is useful for drug discovery and repurposing. These conclusions suggest a model for how a small molecule with few atoms derives selectivity and specificity for particular indications, and to modulate functions of biological systems. Our answers are obtained by exploring the evolutionary nature of the biomolecular interactome (interaction between and within small molecules, proteins, DNA, RNA, etc.). The nature of predicted polypharmacology of small molecule interactions with multiple proteins in the context of an indication/disease illustrates the limitations of the traditional model for drug discovery, and indicates that the interactome based paradigm as implemented by CANDO is a useful model for effective drug discovery and repurposing.

**Most and Least Polypharmacological Drugs Target Different Indications**

The nature of polypharmacology as it applies to drug discovery and repurposing is complex and not a simple function of multitargeting as explained in the previous sections. In order to understand the effect of multitargeting and polypharmacology on drug repurposing, we examine the indications/diseases that are treated by the most and least multitargeting drugs and draw comparisons with imatinib, one of the few well-known polypharmacological drugs (Fig. 5). The five most multitargeting drugs in our study include desmopressin, everolimus, goserelin, buserelin, and cephalothin, which are predicted to strongly interact with over 150 proteins. These drugs are known to treat several complex indications/diseases with multiple etiologies (Fig. 5a), including reducing...
The traditional process of drug discovery is focused on individual compound-protein interactions and requires many rounds of screening, modeling, and synthesis in a trial-and-error approach that is costly, time-consuming, and often ineffective and unsafe in humans. This traditional trials approach, used to develop almost all the drugs to date, has yielded a 95% failure rate before or during clinical use. Existing drugs with known safety profiles that interact with therapeutic targets can be repurposed and rapidly deployed for use in mono and multi-drug therapies. We developed the CANDO platform to evaluate how all approved and other human ingestible drugs interact with all structures representative of the protein universe, and use the resulting information to make predictions of drug efficacy against all indications in a shotgun manner. This approach is highly efficient, produces significantly more drug leads per computing cycle than more conventional methodologies by taking advantage of statistical multiplier effects in much the same manner seen in whole genome shotgun sequencing. The CANDO platform uses signatures of interactions across a representative set of protein structures and determines similarity of compound/drug behavior at a proteomic level (a form of homology modeling). The signatures are used to rank compounds for each indication/disease and provide an optimized and enriched set of compound-protein interactions, a complete and comprehensive list of compounds that may be readily repurposed, as well as a mechanistic understanding of drug behavior at a proteomic level.

The CANDO platform outputs a matrix containing confidence scores measuring the strength of interactions between all human use drugs and a library of protein structures we have constructed as a representation of the protein universe. This evolutionary based drug discovery platform can also be used to predict the small molecule/protein interactions that are implicated in specific indications/diseases, as well as specific biological functions. The CANDO platform can also be used to identify sets of compounds that are different from commonly used drugs in the same pathway. The interaction network information can also be used to determine whether different components act antagonistically or synergistically to guide experiments and identify multiple mechanisms of the same drug.

These structural and evolutionary models of drug discovery are largely static in nature, and are designed for general purpose applications. We can use network-driven methods that integrate heterogeneous temporal and spatial genomics data to specify predictive signatures that identify new candidate therapeutic targets and drugs in a dynamic fashion. This may be done by formulating the problem as one of statistical model selection in which subsets of variables are chosen to build robust models, and adopt the Bayesian paradigm to coherently and rigorously combine different data sources. This leads to a paradigm of disease driven drug discovery, where compounds are designed by accounting for interactions with targets, anti-targets, and off-targets, and the relevant polypharmacology in the context of particular indications is used to identify a drug.

Our work provides a molecular framework to contextualize drug discovery with a paradigm superior to hit-and-trial methodologies currently used by traditional approaches. Altogether, these efforts by us and others to fully quantify the nature and extent of multitargeting/ polypharmacology in the context of specific indications will hopefully be realized in developing new and repurposed therapies to efficiently combat diseases with multiple etiologies (such as cancers, infectious and autoimmune diseases).

**CONFLICT OF INTEREST**

The authors confirm that this article content has no conflict of interest.

**ACKNOWLEDGEMENTS**

Research reported in this publication was supported by a NIH Director's Pioneer Award (1DP1OD006779) to Ram Samudrala and...
a Clinical and Translational Sciences Award (UL1TR001412) from the National Center for Advancing Translational Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Gaurav Chopra gratefully acknowledges support from a Purdue University startup package and JDRF (3-PDF-2014-205-A-N).

REFERENCES
Exploring Polypharmacology in Drug Discovery and Repurposing


Chopra G. A knowledge-based fragment docking with dynamics protocol for protein-ligand docking, unpublished.

Kone J, Samudrala R, Chopra G. Candock: Computational analytics on evolutionary interaction network based docking. unpublished.


Chopra G. Discovery and verification of a novel broad spectrum herpes virus inhibitor. unpublished.

Chopra G. Systems biology based approach to repurpose FDA human approved drugs for treatment of autoimmune diabetes in NOD mice. unpublished.


