Peer-reviewed version available at Molecules 2019, 24, 167; doi:10.3390/molecules24010167

Article

Identifying protein features responsible for improved drug repurposing accuracies using the CANDO platform: Implications for drug design

William Mangione [1], and Ram Samudrala [1,*]

[1] Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, State University of New York, Buffalo, NY 14203, USA; wmangion@buffalo.edu (WM)

* Correspondence: ram@compbio.org; Tel.: +1-414-367-7267

Abstract: Drug repurposing is a valuable tool for combating the slowing rates of novel therapeutic discovery. The Computational Analysis of Novel Drug Opportunities (CANDO) platform performs shotgun repurposing of 2,030 indications/diseases using 3,733 drugs/compounds to predict interactions with 46,784 proteins and relating them via proteomic interaction signatures. An accuracy is calculated by comparing interaction similarities of drugs approved for the same indications. We performed a unique subset analysis by breaking down the full protein library into smaller subsets and then recombining the best performing subsets into larger supersets. Up to 14% improvement in accuracy is seen upon benchmarking the supersets, representing a 100-1,000 fold reduction in the number of proteins considered relative to the full library. Further analysis revealed that libraries comprised of proteins with more equitably diverse ligand interactions are important for describing compound behavior. Using one of these libraries to generate putative drug candidates against malaria results in more drugs that could be validated in the biomedical literature than the list suggested by the full protein library. Our work elucidates the role of particular protein subsets and corresponding ligand interactions that play a role in drug repurposing, with implications for drug design and machine learning approaches to improve the CANDO platform.

Keywords: drug repurposing; drug repositioning; computational biology; drug discovery; computational pharmacology; malaria; multitargeting; malaria treatment

1. Introduction

Common strategies in drug discovery include forward pharmacology [1] and rational drug design [2]. In the former, a library of compounds is screened, typically in high throughout manner, for certain phenotypic effects *in vitro*. In the latter, compounds are virtually screened against a predetermined biological target and high confidence hits are then assayed for a desired modulation. In both cases, the hits obtained are then assessed for effectiveness *in vivo* and proceed to clinical trials for eventual FDA approval if successful at each stage. This iterative process can cost billions of dollars and up to 15 years per drug [3]. These approaches do not consider the promiscuity of approved drugs [4–6] in the context of indications/diseases within living systems (evidenced by side effects present for all small molecule therapies [7,8]), dooming many novel therapeutics to fail. With the second-leading cause of putative drug attrition being adverse reactions [9], there is great utility in finding new uses for already approved drugs, which is formally known as drug repurposing or repositioning [10–13].

We have developed the Computational Analysis of Novel Drug Opportunities (CANDO) platform [14–16] to address these drug discovery challenges. One fundamental tenet of CANDO is that drugs interact with many different proteins and pathways to rectify disease states, and this promiscuous nature is exploited to relate drugs based on their proteomic signatures [14,17–20]. These signatures are typically determined via virtual molecular docking simulations that are applied to predict compound-protein interactions on a proteomic scale. Using a knowledge base of known

drug-indication approvals/associations, we can identify putative drug repurposing candidates for a particular indication based on the similarity of their proteomic interaction signatures to all other drugs approved for (or associated with) that indication. When a particular indication does not have any approved drug, the library of human use compounds present in CANDO is screened against the tertiary structures of all relevant and tractable proteins obtained by x-ray diffraction or homology modeling from a particular organismal proteome to suggest new treatments that maximize binding to the disease-causing proteins and minimize off-target effects. High-confidence putative drug candidates generated by CANDO using both approaches have been prospectively validated preclinically for a variety of indications, including dengue, dental caries, diabetes, hepatitis B, herpes, lupus, malaria, and tuberculosis, with 58/163 candidates yielding comparable or better therapeutic activity than standard treatments [17,18,21,22].

To date, putative drug candidates generated by CANDO have been based on simple comparison metrics, primarily the root mean square deviation (RMSD) of the binding scores present in a pair of drug-proteome interaction signatures. Our platform is evaluated using a benchmarking method that assesses per indication accuracies based on whether or not other drugs associated with the same indication can be captured within a certain cutoff in terms of similarity to a particular drug approved for that indication. Incorporating machine learning, which is continuing to prove its utility in many aspects of biomedicine [23–25] including drug discovery and repurposing [26,27], into the CANDO platform to increase benchmarking accuracies and therefore its predictive power is of importance. Various algorithms can be incorporated (for example, neural networks, support vector machines, and decision trees), but the well documented issues described by the curse of dimensionality [28,29] will plague any choice in the current (v1.5) implementation of CANDO, especially considering the extremely large number of features ($\approx$ 50,000 proteins) within each compound-proteome interaction signature vector. Given the relatively few training samples (an average of $\approx$ 9 drugs per indication), a machine learning approach to train how drugs interact with proteomes is a much easier task with a vastly reduced set of proteins.

Therefore, in this study we set out to find a reduced set of proteins that can therapeutically characterize compounds as well as using the full protein library, with the eventual goal of utilizing them in future machine learning experiments. Our strategy involves using a brute force or greedy feature selection, where the full protein library was split into smaller subsets that were subsequently benchmarked and ranked according to their performance by different metrics. The best performing subsets were recombined into supersets and benchmarked again, which produced substantial improvement with all the benchmarking metrics. Further analysis of the best performing protein subsets and supersets revealed that those that contained proteins predicted to bind to a more equitably diverse ligand structure distribution were strongly associated with increased benchmarking performance, indicating that drugs approved for human use have a specific range and distribution of protein binding site interactions. In addition, protein supersets optimized for independent compound libraries were cross-tested and were able to reproduce the performance of using the full protein library. This indicates that overtraining on a specific compound library during this iterative procedure was limited in scope, making these protein supersets broadly applicable for characterizing drug behavior. We applied one of these protein libraries to generate putative drug candidates against malaria, and the resulting list had more drugs that could be validated in the biomedical literature than the list suggested by the full protein library, indicating the usefulness of the method.

2. Results

2.1. Benchmarking of generated supersets

Fifty iterations of splitting the 46,784 proteins into 5,848 subsets of 8 were performed, resulting in 292,400 benchmarks. The maximum, minimum, mean, and standard deviation were respectively 11.7%, 6.2%, 9.1%, and 0.5% for average indication accuracy, 20.2%, 12.7%, 16.7%, and 0.8% for

pairwise accuracy, and 548, 398, 477.5, and 15.6 for coverage with each metric following a normal distribution (Figure S1). The mean benchmarking performance for each superset within a given iteration tends to gradually increase as the number of iterations increases (Figure 1). Ranking subsets using coverage is overall the best criterion as it yielded the maximum benchmarking performance for all three metrics. Coverage was the dominant ranking criterion beginning at iteration sixteen as it yielded the maximum average indication accuracy and coverage. Average indication accuracy is overall the second best ranking criterion as evidenced by its very close performance to coverage in the average indication accuracy and pairwise accuracy metrics and being the best metric through iteration fifteen. The maximum values obtained by supersets for average indication accuracy, pairwise accuracy, and coverage, were 14.0%, 23.2%, and 602, which represents a 14%, 7%, and 10% improvement on the control (using all 46784 proteins) values of 12.3%, 21.6%, and 546, respectively.

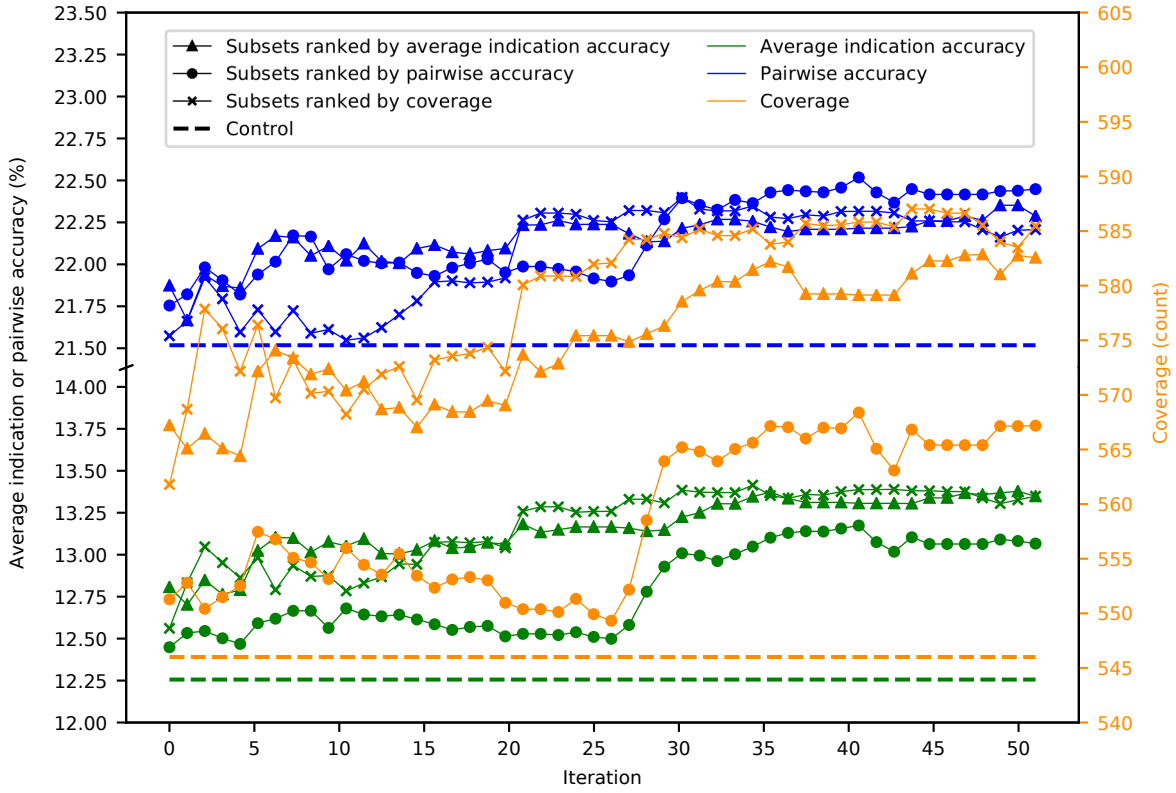

Figure 1. Mean superset benchmarking performance across 50 iterations. Each point represents the mean average indication accuracy (green), pairwise accuracy (blue), or coverage (orange) of 50 supersets created by consecutively combining the top 50 subsets ranked by average indication accuracy (triangles), pairwise accuracy (circles), or coverage (crosses). Average indication accuracy and pairwise accuracy are plotted on the left axis; coverage on the right. The dashed lines represent the control values for their respective metrics. The supersets improve on the control values and gradually increase in performance with the number of iterations with average indication accuracy being the best ranking criterion through ten iterations, after which coverage is superior, especially when measuring the coverage metric. This result demonstrates that the splitting and ranking protocol can produce supersets with benchmarking performance superior to using the full protein library by combining the best performing subsets with a vastly reduced number of proteins (100 to 1,000 fold reduction in size), further suggesting that specific groups of proteins are relatively more useful for drug repurposing accuracy.

Sorting the supersets by size reveals that at least 80-120 proteins are required to reach optimal benchmarking performance (Figure 2). Nonredundantly combining the worst performing subsets of 8

into subpar sets demonstrates worse performance than the control value for the average indication accuracy based on using the full library, with the mean values of these subpar set distributions being below the acceptable 5% threshold of 11.6% for all sizes benchmarked. The random set and subpar set distributions begin to converge toward the average indication accuracy control value as size increases (Figure 2). The principal component analysis (PCA) matrices score very similarly to the random sets, indicating that the supersets are a superior dimensionality reduction method. The mean average indication accuracies begin to plateau or slightly decrease after size 256-264 for the supersets, while continuing to rise as the subpar set size increases. This suggests that the distribution of features within each protein library is important for describing drug/compound behavior.

2.2. Cross-testing with independent compound libraries

For the independent compound library experiment, average indication accuracy was chosen as the ranking criterion because it performed the best in all three metrics through ten iterations in the superset experiment (Figure 1). Box and whisker plots for each compound library show the spread of the benchmarking performance for each metric generated from the protein supersets obtained through the splitting and ranking protocol of their complementary compound library (Figure 3). Supersets less than size 80 were excluded because there is a minimum number of protein features required to reach optimal benchmarking performance (Figure 2). The control value for the given metric fell within the inter-quartile range or below in 26 cases (87%), while it was within the upper quartile in the remaining four cases (13%), indicating that these values always fell within the range of the accuracy/coverage distributions. The control value was an extreme outlier below the distribution in five cases (17%), indicating the potential of these supersets to describe compound behavior more effectively than the full protein library.

2.3. Ligand clustering and feature-based creation of protein libraries

The protein subsets and supersets were analyzed based on four criteria to elucidate the feature(s) responsible for benchmarking performance: organismal source, structure source (x-ray diffraction or modeling), fold space, and interacting ligand structure distributions. There were no significant correlations found for the first three criteria; no organism(s) or fold(s) was consistently represented in the best performing sets, while the the structure sources were evenly distributed among the best and worst performing subsets and supersets (data not shown). All ligands in the COFACTOR database were clustered to investigate the fourth criterion (see Methods). A total of 7,252 ligands were unable to be clustered due to molecular file conversion errors, resulting in 64,592 clustered ligands. Clustering at a distance of 0.3 created 9,929 clusters with varying sizes. The number of clusters including at least 10, 100, and 1,000 compounds were 568, 52, and 7, respectively, including 5,280 singleton clusters. Compound-protein interactions for all 46,784 proteins were mapped to ligand clusters based on the co-crystallized ligand of the binding site that was chosen for each compound-protein pair and ligand cluster signatures were generated for each protein (see Methods). The ligand cluster signatures of the best and worst 50 subsets from the first iteration of the splitting and ranking protocol were averaged together, with the resulting distribution of cluster counts at each rank compared using Welch's T-test (Figure 4). The subsets with the best performance are composed of a much more equitable distribution of interactions among clusters than those with the worst performance on average. The first two ranks show the greatest contrast between the subsets with p-values of $1.04\text{x}10^{-8}$ and $2.38\text{x}10^{-4}$ respectively, with all but two of the 22 rank distributions tested being significantly different (p-value < 0.05). The ligand cluster analysis was repeated for a superset and a subpar set and both follow a similar pattern to the one previously observed, with the superset having a more equitable distribution of interactions among clusters and the subpar set being far more imbalanced.

Based on the subset analysis, it was hypothesized that proteins having a more diverse and equitable distribution of interactions will benchmark better. Protein libraries were created by ranking

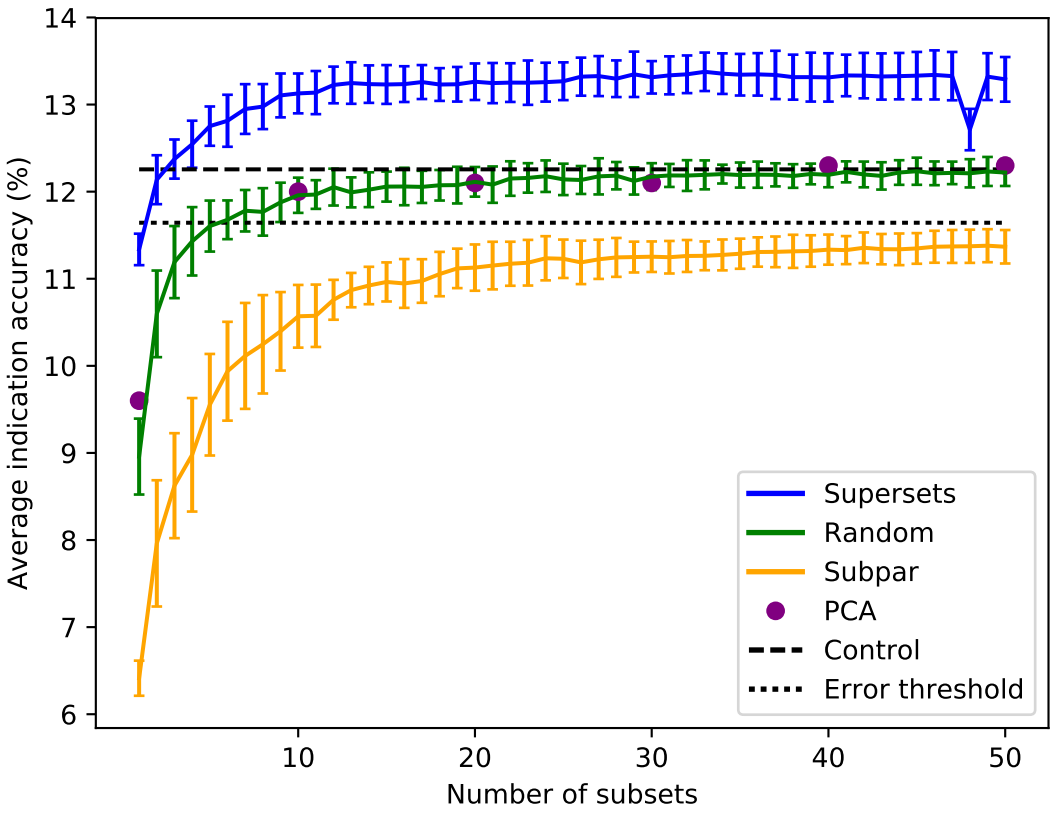

Figure 2. Superset, subpar set, random set, and PCA matrices average indication accuracies sorted by size. Average indication accuracies are shown for the supersets (blue) generated using the best subsets ranked by the same metric. The line traces the mean score for each size with the bars indicating one standard deviation for the distribution. Subpar sets (orange) are the combinations of the worst performing subsets ranked by average indication accuracy. Randomly selected protein sets (green) of each size were also generated and benchmarked. Principal component analysis (PCA, purple circles) was used to reduce the complete compound-protein interaction matrix to 8, 80, 160, 240, 320, and 400 dimensions. The control value based on using the full protein library (dashed black at 12.3%) and an acceptable 5% threshold (dotted black at 11.6%) are plotted for reference (i.e., any protein set that benchmarks within 95% of the control value is considered acceptable). For the random sets and supersets, the performance in terms of average indication accuracy begins to plateau around 80-120 proteins. The supersets begin to slightly decline in performance after 32-33 subsets (256-264 proteins). The mean subpar set accuracies at each size all fall outside the 5% acceptable threshold, while the superset distributions are well above the control value with as few as five subsets. The PCA matrices perform similarly to the control and the random sets. The difference between the superset and subpar set performance suggests that there is a particular distribution of features within their proteins that is correlated with benchmarking performance.

the 46,784 proteins in CANDO on the variance of the number of interactions attributed to each ligand cluster (see Methods). Creating libraries using proteins with the lowest variance of ligand cluster interactions (excluding the trivial case of proteins mapped to only one cluster) results in benchmarking performance much higher than libraries composed of proteins with high variance (Figures 5 and 6). To elucidate the impact of the number of ligand clusters mapped to a given protein, we applied a upper cutoff filter for the number of ligand clusters allowed (Figure 5); all libraries were limited to size 80, which is the minimum required to reach optimal performance based on data in Figure 2. There is

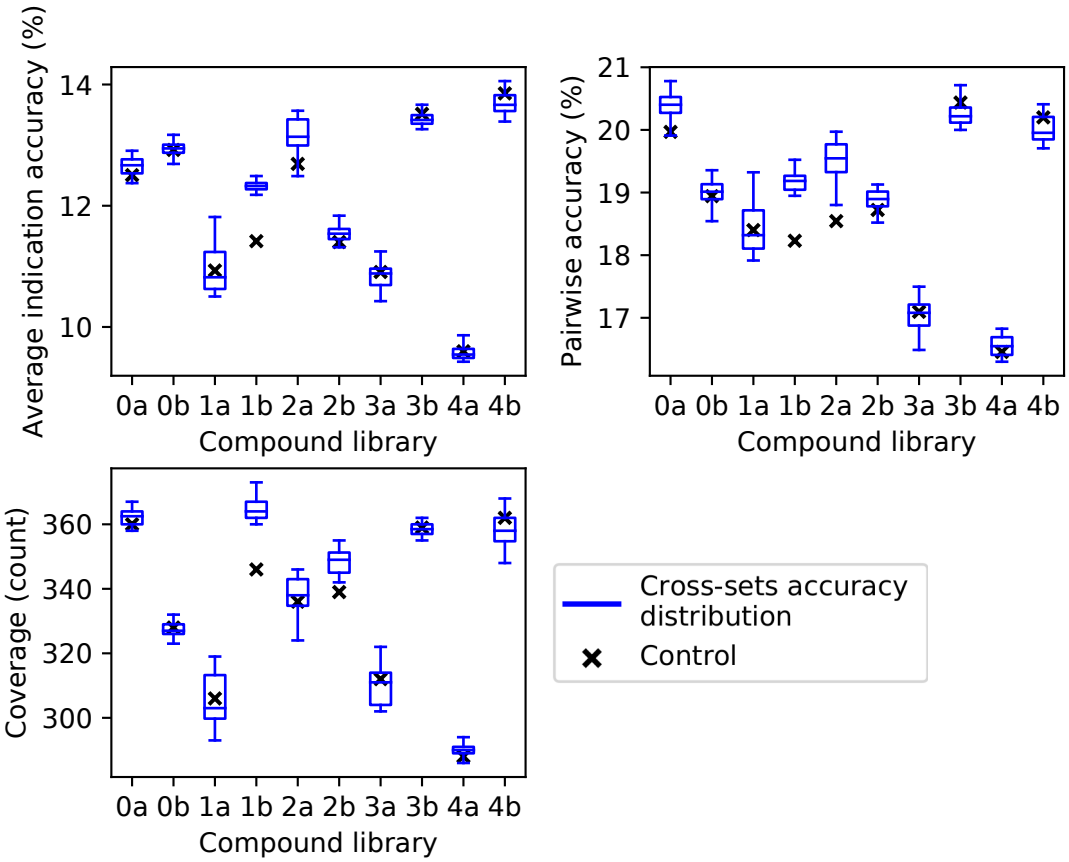

Figure 3. Benchmarking performance of protein supersets cross-tested with independent compound libraries. Protein supersets were generated using the splitting and ranking protocol with the average indication accuracy metric. Supersets were tested on their complementary library only (for instance, supersets generated from compound library 0a were tested on compound library 0b and vice versa). The blue box and whisker plots describe the benchmarking performance distributions of the 41 complementary supersets with the middle line being the median value, the box encompassing the first and third quartiles, and the whiskers extending to the maximum and minimum excluding outliers. The nine supersets less than size 80 were excluded due to results in Figure 2. In all cases, the control value for each compound library (black cross) never lies above the distribution of cross superset accuracies/coverages, indicating that these supersets can be generalized to any library of compounds.

an optimal upper limit of ≈20-40 ligand clusters with regard to the benchmarking, indicating that using too many ligand clusters to describe a protein is undesirable for characterizing drug/compound behavior. Based on the data in Figure 5, we created protein libraries of various sizes using a upper cutoff of 40 (Figure 6). We begin to consistently recapture the benchmarking performance of the full library (within 5% error) with as few as 60 proteins. Libraries composed of high variance proteins, which are proteins with greatly imbalanced ligand cluster signatures (i.e. $>$ 95% interactions mapped to one cluster), produce benchmarking performance outside of the acceptable 5% error range for all the created libraries (Figures 5 and 6). Figure 7 provides a visual depiction of the best and worst types of proteins for benchmarking performance, highlighting the idea that a moderately diverse and equitable distribution of interacting ligand clusters is ideal for describing drug behavior.

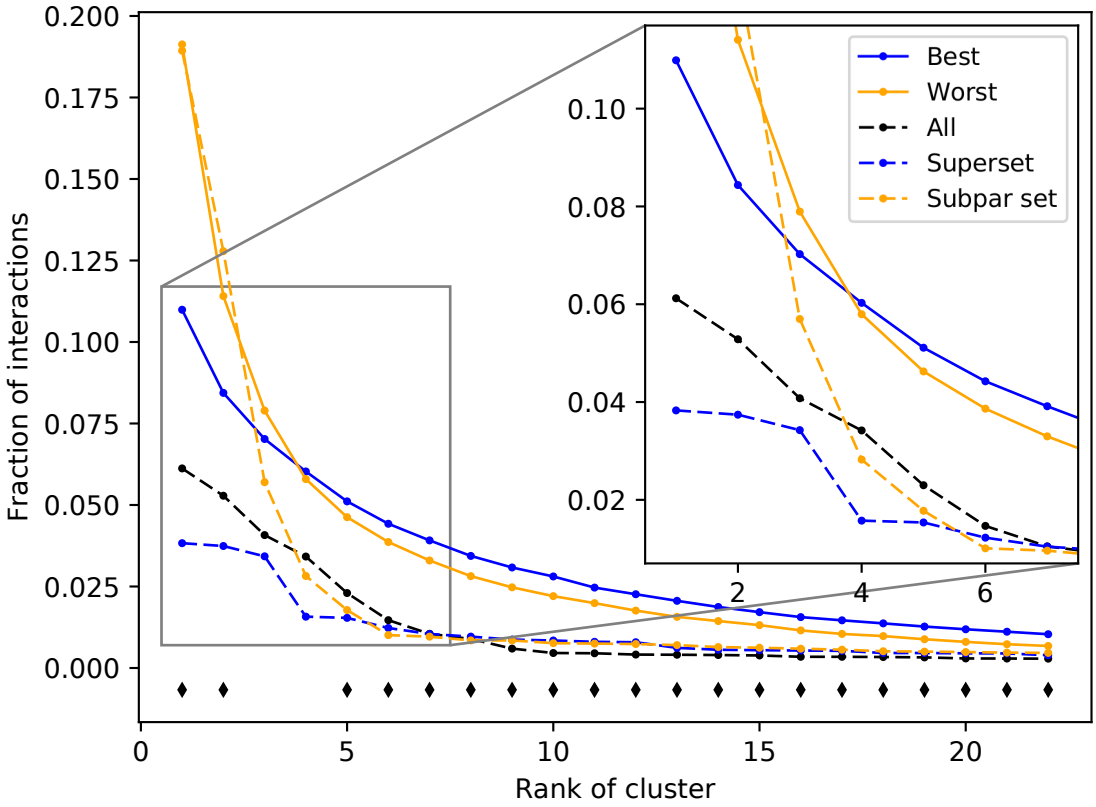

Figure 4. Fraction of interactions attributed to top ranked ligand clusters from different protein sets. The solid blue and solid orange points are averages of 50 best and worst subsets, respectively. Dashed lines represent an example of a superset (blue) and a subpar set (orange) which are nonredundant combinations of the best and worst performing subsets respectively. The control set (dashed black), representing the full protein structure library, falls in between the superset and subpar set. The black diamonds indicate that the distribution of counts at that cluster rank between the best and worst performing subsets, assessed using Welch's T-test, is significant (p-value < 0.05). The subsets and supersets with the best performance demonstrate a more equitable distribution of interactions among ligand clusters as opposed to the worst performing subsets and subpar sets, indicating that using multitargeting proteins to compose our structure libraries yields superior benchmarking performance.

2.4. Validation case study: malaria

The lists of the putative drug candidates against malaria generated using both the created protein library and the full protein library are available in the Supplementary Information. Many candidates are shared between the two libraries, including closely related compounds quinine ethyl carbon-ate and cinchonine monohydrochloride, as well as tigecycline, with the former two having known anti-malarial activity and the latter having success *in vitro* and *in vivo* [31,32]. A trivial candidate predicted using the full library is quinacrine, a malaria treatment closely related to and since supplanted by the more efficacious chloroquine. The top two candidates exclusively generated using the created protein library are the antifungal drug posaconazole and experimental compound zosuquidar, both of which have shown anti-malarial activity as lactate dehydrogenase enzyme and P-glycoprotein specific inhibitors, respectively [33,34]. None of the top candidates generated using the full protein library that were not also suggested using the created protein library could be found to have activity against malaria in our search of the biomedical literature.

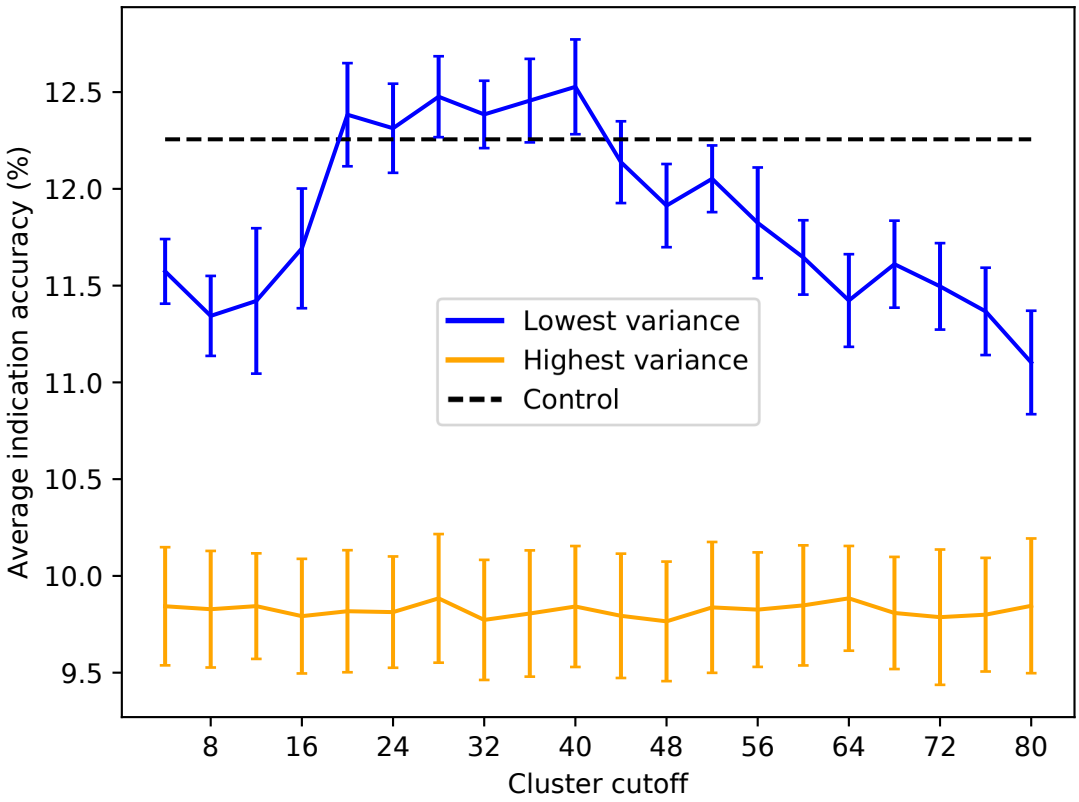

Figure 5. Average indication accuracy performance of created protein libraries of size 80 with various upper ligand cluster cutoffs. Protein libraries were created by randomly selecting 80 of the top 160 proteins from the list of proteins ranked by ligand cluster interaction variance (see Methods). A minimum of two ligand clusters was required to avoid the trivial case of only one cluster mapped (with a variance of zero). A upper ligand cluster count cutoff was applied to these protein libraries to determine the effect on benchmarking performance. The blue line traces the average indication accuracy distribution mean from benchmarking 50 libraries at each cutoff using the top ranked proteins. Similarly, the orange line traces the average indication accuracy distribution mean from benchmarking 50 libraries at each cutoff using the highest variance proteins. The bars indicate one standard deviation for the distribution. Using too small ($<$ 20) or too large ($>$ 40) of a cutoff results in suboptimal benchmarking performance. The high variance libraries result in average performance far below the acceptable 5% error range, with the cluster cutoff seemingly having no effect on average indication accuracy as all cutoffs produced comparable distributions. All average indication accuracies produced using a upper cutoff of 20-40 ligand clusters were within the acceptable 5% error range, with the upper cutoff of 40 ligand clusters producing the greatest mean accuracy. This result indicates that there is an optimal range of ligand cluster interactions to best describe therapeutic behavior.

3. Discussion

The splitting and ranking protocol was originally intended to find a protein subset that benchmarked as least as well as the full set. The improvement of the benchmarking performance is an encouraging sign for incorporating machine learning in the CANDO platform in the future, and discovering how more complex weighting and relating of proteins contribute to drug repurposing accuracy, which is difficult to do with simple RMSD calculations. The smaller-sized protein libraries generated as part of this study, representing a 100 to 1,000 fold reduction in size, will be more conducive

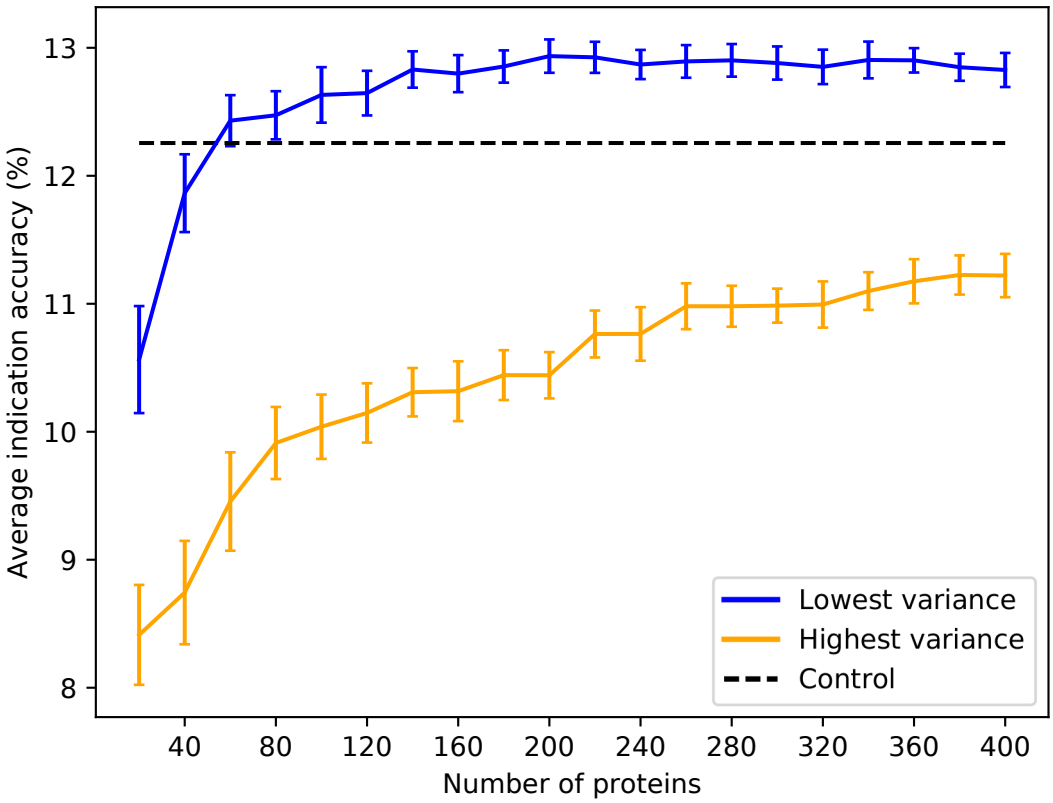

Figure 6. Average indication accuracy performance of created protein libraries at various sizes with ligand cluster cutoff of 40. Protein libraries were created by randomly selecting half the number of proteins in the library ranked by ligand cluster interaction variance with 2 to 40 ligand clusters mapped. The blue line traces the average indication accuracy distribution mean from benchmarking 50 libraries at each size using the top ranked proteins. Similarly, the orange line traces the average indication accuracy distribution mean from benchmarking 50 libraries at each size using the highest variance proteins. The bars indicate one standard deviation for the distribution. Using too small of a size (< 60) results in suboptimal benchmarking performance. Creating libraries from proteins with the highest variance results in performance on average far below the acceptable 5% error range, although size does have a positive correlation with performance with these high variance sets. This result reiterates that there is a minimum number of proteins required to reach optimal benchmarking performance and that proteins with high variance in their ligand cluster signatures are far superior for describing drug/compound behavior.

to machine learning. Feature reduction through approaches other than PCA, such as neural network based auto-encoders, will provide an important contrast to our proposed method.

The independent compound library experiment demonstrated that optimized protein sets based on a particular library were capable of therapeutically characterizing a completely different one, indicating that these supersets are generalizable. In other words, if a new drug/compound is added to the CANDO putative drug library, these reduced size supersets are likely able to describe its behavior at least as well as using every protein available. In addition to facilitating machine learning, our findings suggest a greatly reduced time required to generate new proteomic interaction vectors, which is particularly important if the program/protocol of choice for generating interactions is computationally expensive. Any repurposing candidates suggested from using the supersets are on average more

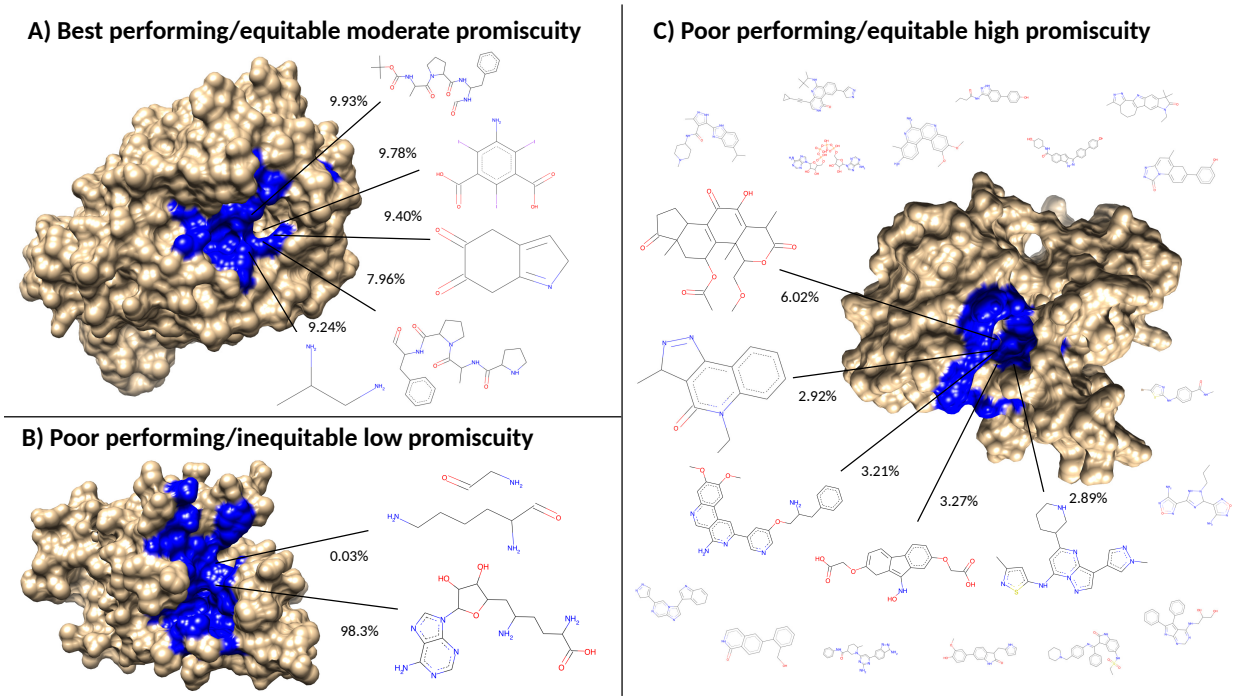

Figure 7. Visualization of the best and worst protein types for benchmarking performance based on ligand cluster signatures. Centroids of the top five ligand clusters from the signatures of each protein are depicted. The percent of interactions belonging to each ligand cluster are next to their respective centroid. Surface representations of the proteins were made using Chimera [30] with the predicted binding site residues from COFACTOR for each ligand shown (excluding the smaller ligands in C) colored in blue. A) Alkaline serine protease KP-43 from *Bacillus subtillus*: the top five ligand clusters account for 46.3% of the total interactions with the distribution between them being relatively equitable. B) SET domain of human histone-lysine N-methyltransferase: only two ligand clusters are predicted to interact, with one having over 98% of the total interactions. C) Human STE20-related kinase adapter protein beta: the ligand cluster signature is too promiscuous with the top five ligand clusters accounting for only 18.3% of the total interactions; the remaining sixteen ligands surrounding the protein account for 28.7%, which combined with the top five clusters is as much as the 46.3% of the total interactions shown in A from only five clusters. Subsets and supersets composed of proteins similar to A outperform those composed of proteins similar to B and C in benchmarking, indicating that moderately promiscuous proteins with equitable ligand cluster signatures are the best therapeutic descriptors.

clinically relevant as they were able to recapture drug behavior more accurately than using the full protein library in a statistically significant manner.

The knowledge that ligand cluster interaction diversity is the key requisite feature in describing drug behavior may also lead to optimization strategies other than the computationally intensive splitting and ranking protocol used in this study. The importance of ligand cluster interaction diversity may play a role in a variety of applications in systems and synthetic biology, for example in the design of protein systems that are specifically tailored to handle drug absorption, distribution, metabolism, and excretion processes. These optimizations incorporated into our repurposing platform should result in the rapid generation of more accurate putative drug repurposing candidates, thereby alleviating the problems associated with current drug discovery. To demonstrate this in a cursory fashion, we used a specific reduced size library of 100 proteins to generate putative drug candidates against malaria, which captured more candidates with known anti-malarial activity compared to using the full library.

The ligand cluster analysis revealed that compound-protein interactions are more therapeutically relevant if the proteins used to describe the behavior of a compound are diverse in terms of the

structures of the ligands which interact with their binding sites. Protein libraries with fewer predicted ligand cluster binding partners yield much worse performance than those consisting of proteins interacting with a more structurally diverse range of ligands. Coupling this with the finding that there is a minimum number of proteins required to reach optimal benchmarking accuracies (Figure 2), which was also observed by us previously [15,16], drugs should realistically be described in the context of their multitarget nature, treating both small molecule compounds and proteins promiscuously, as in biological systems [4,35,36]. However, using libraries of proteins with too many diverse interactions in the CANDO platform also leads to suboptimal performance. We hypothesize this can be attributed to two factors: 1) spreading a compound interaction signature across too many (50 or more) ligand clusters can potentially dilute the therapeutic signal relative to the promiscuity of the corresponding proteins; or 2) these proteins are not therapeutically relevant and are therefore not useful for specifically describing drug behavior.

In the context of drug design, comprehending the promiscuous nature of small molecule drugs to develop strategies that optimize their interactions with macromolecular targets of interest, while minimizing interactions leading to adverse events, will address the two leading causes of drug attrition in clinical trials [9]. To this end, further studies to enhance the CANDO platform using a variety of molecular fingerprint descriptors for compounds, docking methods, compound-protein interaction parameters, machine learning methods, side effect data, and indication-specific supersets, are currently underway.

4. Materials and Methods

4.1. Compound and indication mappings

The putative drug library in the CANDO platform comprises 3,733 human use compounds, including all clinically approved drugs from the US FDA, Europe, Canada, and Japan, collected from DrugBank [37], NCGC Pharmaceutical Collection (NPC) [38], Wikipedia, and PubChem [39]. Each drug/compound was converted to a three dimensional (3D) or tertiary structure to standardize input conformation and avoid biasing the results using ChemAxon's MarvinBeans molconverter v.5.11.3 (https://chemaxon.com/). InChiKeys were generated from all preprocessed compounds using Xemistry's Cactvs Chemoinformatics Toolkit [40] (https://www.xemistry.com/) to remove redundancies. Drug-disease associations were obtained from the Comparative Toxicogenomics Database (CTD) [41] and mapped to our drug/compound library, resulting in associations to 2,030 indications, including 1,439 indications with at least two drugs that are used to perform our leave-one-out benchmarking protocol described below [14–16,18,20].

4.2. Compound-proteome interaction matrix generation

The protein structure library in the CANDO platform is made up of 48,278 tertiary conformations solved using x-ray diffraction taken from the Protein Data Bank (PDB) [42] (≈80% of the structures) as well as homology models (≈20%). The organism sources for the proteins include *Homo sapiens*, and several higher order eukaryotes, bacteria and viruses. Protein structure models were generated using HHBLITS [43], I-TASSER [44,45], and KoBaMIN [46]. KoBaMIN uses knowledge-based force fields for fast protein model structure refinement, while ModRefiner [45] also uses physics based force fields for the same purpose. HHBLITS uses hidden Markov models to increase speed and accuracy of protein sequence alignments and LOMETS [47] uses multiple threading programs to align and score protein targets and templates. SPICKER [48] identifies native protein folds by clustering the computer generated models. The I-TASSER modeling pipeline consists of the following steps: 1) HHBLITS and LOMETS for template model selection; 2) threading of protein sequences from templates as structural fragments; 3) replica-exchange Monte Carlo simulations for fragment assembly; 4) SPICKER for clustering of simulation decoys; 5) ModRefiner for generation of atomically refined model SPICKER centroids; 6) KobaMIN for final refinement of models. Some pathogen proteins

failed during the modeling and were removed, ultimately resulting in 46,784 proteins in the final matrix. To generate scores for each compound-protein interaction, COFACTOR [49] was first used to determine potential ligand binding sites for each protein by scanning a library of experimentally determined template binding sites with the bound ligand from the PDB. COFACTOR outputs multiple binding site predictions, each with an associated binding site score. For each predicted binding site, the associated co-crystallized ligand is compared to each compound in our set using the OpenBabel FP4 fingerprinting method [50], which assesses compound similarity based on functional groups from a set of SMARTS [51] patterns, resulting in a structural similarity score. The score that populates each cell in the compound-protein interaction matrix is the maximum value of all of the possible binding site scores times the structural similarity scores of the associated ligand and the compound.

4.3. Benchmarking protocol and evaluation metrics

The compound-compound similarity matrix is generated using the root mean square deviation (RMSD) calculated between every pair of compound interaction signatures (the vector of 46,784 real-value interaction scores between a given compound and every protein in the library). Two compounds with a low RMSD value are hypothesized to have similar behavior [14–16,18,20]. For each of the 1,439 indications with two or more associated drugs, the leave-one-out benchmark assesses accuracies based on whether another drug associated with the same indication can be captured within a certain cutoff of the ranked compound similarity list of the left-out drug. This study primarily focused on a cutoff of the ten most similar compounds (“top10"), the most stringent cutoff used in previous publications [14–16,18,20]. The benchmarking protocol calculates three metrics to evaluate performance: average indication accuracy, compound-indication pairwise accuracy, and coverage. Average indication accuracy is calculated by averaging the accuracies for all 1439 indications using the formula c/d x 100, where c is the number of times at least one drug was captured within the cutoff (top10 in this study) and d is the number of drugs approved for that given indication. Pairwise accuracy is the weighted average of the per indication accuracies based on how many drugs are approved for a given indication. Coverage is the count of the number of indications with non-zero accuracies within the top10 cutoff.

4.4. Superset creation and benchmarking

The 46,784 proteins in the CANDO platform were randomly split into 5,848 subsets of 8 and subsequently benchmarked using the method described above. The size of 8 was selected because offered the widest range of benchmarking values (relative to larger sizes), reduced the computational cost of the experiments (relative to smaller sizes that increase the number of individual benchmarks that need to be evaluated), divided into 46,784 evenly, and also provided adequate signal for the multitargeting approach to work according to our prior studies [17]. A total of 50 iterations were performed that resulted in 292,400 benchmarking experiments. Each subset was then ranked according to top10 average indication accuracy, pairwise accuracy, and coverage. The fifty best performing subsets from each ranking criterion (average indication accuracy, pairwise accuracy, and coverage) were progressively combined into supersets and benchmarked after each of the 50 iterations of the splitting and ranking protocol. The subsets were nonredundantly combined such that if a given protein was represented in the best performing subsets more than once (from two or more different iterations), then it would only occur once in the superset. The number of proteins in each superset ranged from 8 to 400. The complete protein-compound interaction matrix was reduced to 8, 80, 160, 240, 320, and 400 dimensions using principal component analysis (PCA) to serve as a control.

4.5. Independent compound library creation and evaluation

The CANDO putative drug library was split into two disjoint libraries comprising ≈50% of the compounds (1,866 and 1,867 compounds) five times. For each such library, the indication mapping was reconstructed to include only the disease associations of the drugs present. Each corresponding

pipeline comprised of a disjoint compound library was subjected to ten iterations of the splitting and ranking protocol. Protein supersets, composed of the fifty best performing subsets for each metric that were progressively combined (see previous section), were generated and benchmarked. Supersets less than size 80 were excluded due to the results of Figure 2, which shows that benchmarking performance begins to plateau starting with around that number of proteins. The remaining supersets were then cross-tested on their complementary sets to assess the generalizability of our selection and optimization protocol by comparing it to the corresponding control value based on using the full protein structure library (but with the same disjoint compound library). The benchmarking for each of these 50% disjoint compound libraries are not directly comparable to each other nor to the full compound library because the indication mappings are different.

4.6. Evaluating the features responsible for protein subset and superset accuracy

The best performing protein subsets and supersets were further analyzed to elucidate the protein feature(s) responsible for increased benchmarking performance. The protein subsets and supersets were analyzed based on four criteria: organismal source, structure source (x-ray diffraction or modeling), fold space (based on the CATH classification of proteins [52]), and interacting ligand structure distributions. The subsets and supersets were analyzed by counting the specific organisms to which the proteins belonged to see if any were over or underrepresented in the best and worst performing sets. Similarly, the subsets and supersets were analyzed to see if structures obtained via a specific source, x-ray diffraction or modeling, were differentially represented in the best and worst performing sets. Fold assignments were made to each protein in the subsets and supersets which were again analyzed for differential representation of specific protein folds. Finally, since our compound-protein interaction scoring method utilized the structural similarity of each drug/compound to the ligand co-crystallized with the protein (see previous section), we analyzed these ligands for differential representation. Each co-crystallized ligand in the COFACTOR database of template binding sites were clustered at various distances (0.1 to 0.9 with increments of 0.1) using the Butina clustering algorithm [53] in the RDKit [54] library based on the Tanimoto distance [55] of their Daylight fingerprints (http://www.daylight.com/). A total of 64,592 ligands were clustered with 7,252 ligands failing due to molecular file conversion errors. Each protein in the subsets and supersets was assigned a ligand cluster signature where each value in the vector is the number of times a ligand from a given cluster was chosen while calculating the compound-protein interaction score protocol for that protein. The fraction of interactions belonging to each cluster for each protein set was calculated by first adding the ligand cluster signatures for each protein belonging to the set, then ranking the clusters from greatest to smallest, and finally dividing by the number of total interactions (size of the protein set times 3,733). The ligand cluster signatures of the best and worst 50 subsets from the first iteration of the splitting and ranking protocol were averaged together, with the resulting distribution of cluster counts at each rank compared using Welch's T-test for equal means. Protein subsets were chosen from only one iteration to ensure independence.

4.7. Creating protein libraries using the most important feature

Of the features evaluated (see above), only the ligand cluster distribution could be correlated with benchmarking performance (see Results). We then generated new protein structure libraries that captured this feature ideally and assessed benchmarking performance. Each protein was ranked based on the variance of the values in their ligand cluster signatures. A minimum cutoff of two clusters was used to prevent the trivial case of proteins with only one cluster mapped (with a variance of zero). Another cutoff of at least 1,867 total mapped interactions was used to account for proteins with interactions mapped to unclustered ligands. Proteins were randomly selected from the top of the ranked list of variances at size increments of 20 and then benchmarked. The cutoff of proteins considered for random selection was two times the size of the library (for instance, 100 proteins were randomly selected from the top 200 proteins ranked by ligand cluster signature variance). A total of 50

libraries were made for each size to generate a distribution of benchmarking values. This procedure was repeated for the bottom of the ranked list of variances for comparison.

4.8. Validation case study: malaria

We analyzed the ability of the created protein libraries to generate putative drug candidates for the treatment of malaria, an indication targeted in previous publications [16,17]. We used all the indications belonging to the MeSH categories related to malaria in our drug-indication mapping, namely "malaria" (MESH:D008288) "falciparum malaria" (MESH:D016778), "vivax malaria" and "cerebral malaria" (MESH:D016779), and a library of 100 proteins from the previous step that achieved a top10 benchmarking accuracy of 13.3%, to generate drug/compound similarities using our platform. We then used a concurrence-ratio scoring method to generate candidates by first counting the number of times each compound appears in the top 10 most similar compounds of each drug approved for a malaria indication and then ranking the compounds by dividing by the number of drugs approved for that malaria indication. We compared this to the putative drug candidate list generated by the full library of 46,784 proteins and searched the biomedical literature for validation using Google Scholar.

5. Conclusions

We have developed an integrated pipeline that allows for the elucidation of proteins and their features that are important for benchmarking in the CANDO platform, and therefore important for drug repurposing. We are able to reproduce the performance of the complete CANDO protein structure library with orders of magnitude fewer proteins, allowing for more rapid candidate generation when evaluating new putative drug libraries or any other changes to the platform. We discovered that moderately promiscuous proteins, in terms of the structures of ligands with which they are predicted to interact, are important for describing how drugs behave in biological systems, a claim validated by literature evidence supporting putative drug candidates generated by a library composed of a subset of these proteins for the treatment of malaria. The implications for drug design are that appreciating the multitarget nature of small molecule therapies and optimizing their interactions with the range of macromolecular targets that they are exposed to in their environments during their absorption, dispersion, metabolism, and excretion may be more fruitful than traditional rational drug design using single targets.

Supplementary Materials: The following are available online, Figure S1: Distributions of protein subset benchmarking accuracies and coverages, File S1: Putative drug candidates against malaria generated using the full protein library, File S2: Putative drug candidates against malaria generated using a created library of 100 proteins, and File S3: Drugs approved for the treatment of malaria used in this study.

Author Contributions: William Mangione conceptualized the experiments, designed the methodology, performed all formal analysis, and drafted the manuscript. Ram Samudrala provided general oversight and mentorship, helped with data interpretation, and helped with manuscript development and editing.

Funding: This work was supported in part by a 2010 National Institute of Health Director's Pioneer Award [1DP1OD006779], a National Institute of Health Clinical and Translational Sciences Award [UL1TR001412], a National Library of Medicine T15 Award [T15LM012495], a NCI/VA BD-STEP Fellowship in Big Data Sciences, and startup funds from the Department of Biomedical Informatics at the University at Buffalo.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Broach, J.R.; Thorner, J. High-throughput screening for drug discovery. *Nature* **1996**, *384*, 14–16.
2. Macalino, S.J.Y.; Gosu, V.; Hong, S.; Choi, S. Role of computer-aided drug design in modern drug discovery. *Archives of pharmacal research* **2015**, *38*, 1686–1701.
3. Mullard, A. New drugs cost US [dollar] 2.6 billion to develop. *Nature Reviews Drug Discovery* **2014**, *13*, 877–877.

4. L Bolognesi, M. Polypharmacology in a single drug: multitarget drugs. *Current medicinal chemistry* **2013**, *20*, 1639–1645.
5. Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: challenges and opportunities in drug discovery: miniperspective. *Journal of medicinal chemistry* **2014**, *57*, 7874–7887.
6. Hu, Y.; Bajorath, J. Monitoring drug promiscuity over time. *F1000Research* **2014**, *3*.
7. Iwata, H.; Mizutani, S.; Tabei, Y.; Kotera, M.; Goto, S.; Yamanishi, Y. Inferring protein domains associated with drug side effects based on drug-target interaction network. *BMC systems biology* **2013**, *7*, S18.
8. Liu, T.; Altman, R.B. Relating essential proteins to drug side-effects using canonical component analysis: a structure-based approach. *Journal of chemical information and modeling* **2015**, *55*, 1483–1494.
9. Arrowsmith, J.; Miller, P. Trial watch: phase II and phase III attrition rates 2011-2012. *Nature Reviews Drug Discovery* **2013**, *12*, 569–569.
10. Liu, X.; Zhu, F.; H Ma, X.; Shi, Z.; Y Yang, S.; Q Wei, Y.; Z Chen, Y. Predicting targeted polypharmacology for drug repositioning and multi-target drug discovery. *Current medicinal chemistry* **2013**, *20*, 1646–1661.
11. Achenbach, J.; Tiikkainen, P.; Franke, L.; Proschak, E. Computational tools for polypharmacology and repurposing. *Future medicinal chemistry* **2011**, *3*, 961–968.
12. Yella, J.; Yaddanapudi, S.; Wang, Y.; Jegga, A. Changing trends in computational drug repositioning. *Pharmaceuticals* **2018**, *11*, 57.
13. Hurle, M.; Yang, L.; Xie, Q.; Rajpal, D.; Sanseau, P.; Agarwal, P. Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics* **2013**, *93*, 335–341.
14. Chopra, G.; Samudrala, R. Exploring polypharmacology in drug discovery and repurposing using the CANDO platform. *Current pharmaceutical design* **2016**, *22*, 3109–3123.
15. Sethi, G.; Chopra, G.; Samudrala, R. Multiscale modelling of relationships between protein classes and drug behavior across all diseases using the CANDO platform. *Mini reviews in medicinal chemistry* **2015**, *15*, 705–717.
16. Minie, M.; Chopra, G.; Sethi, G.; Horst, J.; White, G.; Roy, A.; Hatti, K.; Samudrala, R. CANDO and the infinite drug discovery frontier. *Drug discovery today* **2014**, *19*, 1353–1363.
17. Jenwitheesuk, E.; Samudrala, R. Identification of potential multitarget antimalarial drugs. *JAMA* **2005**, *294*, 1487–1491.
18. Chopra, G.; Kaushik, S.; Elkin, P.L.; Samudrala, R. Combating ebola with repurposed therapeutics using the CANDO platform. *Molecules* **2016**, *21*, 1537.
19. Jenwitheesuk, E.; Horst, J.A.; Rivas, K.L.; Van Voorhis, W.C.; Samudrala, R. Novel paradigms for drug discovery: computational multitarget screening. *Trends in pharmacological sciences* **2008**, *29*, 62–71.
20. Horst, J.A.; Laurenzi, A.; Bernard, B.; Samudrala, R. Computational multitarget drug discovery. *Polypharmacology* **2012**, pp. 236–302.
21. Horst, J.; Pieper, U.; Sali, A.; Zhan, L.; Chopra, G.; Samudrala, R.; Featherstone, J. Strategic protein target analysis for developing drugs to stop dental caries. *Advances in dental research* **2012**, *24*, 86–93.
22. Costin, J.M.; Jenwitheesuk, E.; Lok, S.M.; Hunsperger, E.; Conrads, K.A.; Fontaine, K.A.; Rees, C.R.; Rossmann, M.G.; Isern, S.; Samudrala, R. Structural optimization and de novo design of dengue virus entry inhibitory peptides. *PLoS neglected tropical diseases* **2010**, *4*, e721.
23. Obermeyer, Z.; Emanuel, E.J. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine* **2016**, *375*, 1216.
24. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* **2015**, *13*, 8–17.
25. Shen, D.; Wu, G.; Suk, H.I. Deep learning in medical image analysis. *Annual review of biomedical engineering* **2017**, *19*, 221–248.
26. Menden, M.P.; Iorio, F.; Garnett, M.; McDermott, U.; Benes, C.H.; Ballester, P.J.; Saez-Rodriguez, J. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one* **2013**, *8*, e61318.
27. Bakheet, T.M.; Doig, A.J. Properties and identification of human protein drug targets. *Bioinformatics* **2009**, *25*, 451–457.
28. Keogh, E.; Mueen, A., Curse of dimensionality. In *Encyclopedia of Machine Learning and Data Mining*; Springer, 2017; pp. 314–315.

29. Friedman, J.H. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery* **1997**, *1*, 55–77.
30. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* **2004**, *25*, 1605–1612.
31. Sahu, R.; Walker, L.A.; Tekwani, B.L. In vitro and in vivo anti-malarial activity of tigecycline, a glycylcycline antibiotic, in combination with chloroquine. *Malaria journal* **2014**, *13*, 414.
32. Starzengruber, P.; Thriemer, K.; Haque, R.; Khan, W.; Fuehrer, H.; Siedl, A.; Hofecker, V.; Ley, B.; Wernsdorfer, W.; Noedl, H. Antimalarial activity of tigecycline, a novel glycylcycline antibiotic. *Antimicrobial agents and chemotherapy* **2009**, *53*, 4040–4042.
33. Alcantara, L.M.; Kim, J.; Moraes, C.B.; Franco, C.H.; Franzoi, K.D.; Lee, S.; Freitas-Junior, L.H.; Ayong, L.S. Chemosensitization potential of P-glycoprotein inhibitors in malaria parasites. *Experimental parasitology* **2013**, *134*, 235–243.
34. Penna-Coutinho, J.; Cortopassi, W.A.; Oliveira, A.A.; França, T.C.C.; Krettli, A.U. Antimalarial activity of potential inhibitors of Plasmodium falciparum lactate dehydrogenase enzyme selected by docking studies. *PloS one* **2011**, *6*, e21237.
35. Hopkins, A.L. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology* **2008**, *4*, 682.
36. Hart, T.; Dider, S.; Han, W.; Xu, H.; Zhao, Z.; Xie, L. Toward repurposing metformin as a precision anti-cancer therapy using structural systems pharmacology. *Scientific reports* **2016**, *6*, 20441.
37. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A.C.; Wishart, D.S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* **2011**, *39*, D1035–41. doi:10.1093/nar/gkq1126.
38. Huang, R.; Southall, N.; Wang, Y.; Yasgar, A.; Shinn, P.; Jadhav, A.; Nguyen, D.T.; Austin, C.P. The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci Transl Med* **2011**, *3*, 80ps16. doi:10.1126/scitranslmed.3001862.
39. Li, Q.; Cheng, T.; Wang, Y.; Bryant, S.H. PubChem as a public resource for drug discovery. *Drug Discov Today* **2010**, *15*, 1052–7. doi:10.1016/j.drudis.2010.10.003.
40. Ihlenfeldt, W.D.; Takahashi, Y.; Abe, H.; Sasaki, S.i. Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and compatibility. *Journal of chemical information and computer sciences* **1994**, *34*, 109–116.
41. Davis, A.P.; Murphy, C.G.; Johnson, R.; Lay, J.M.; Lennon-Hopkins, K.; Saraceni-Richards, C.; Sciaky, D.; King, B.L.; Rosenstein, M.C.; Wiegers, T.C. The comparative toxicogenomics database: update 2013. *Nucleic acids research* **2012**, *41*, D1104–D1114.
42. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E., The protein data bank, 1999–. In *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*; Springer, 2006; pp. 675–684.
43. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* **2012**, *9*, 173.
44. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics* **2008**, *9*, 40.
45. Xu, D.; Zhang, J.; Roy, A.; Zhang, Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins: Structure, Function, and Bioinformatics* **2011**, *79*, 147–160.
46. Rodrigues, J.P.; Levitt, M.; Chopra, G. KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic acids research* **2012**, *40*, W323–W328.
47. Wu, S.; Zhang, Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research* **2007**, *35*, 3375–3382.
48. Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **2004**, *57*, 702–710.
49. Roy, A.; Yang, J.; Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research* **2012**, *40*, W471–W477.
50. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *Journal of cheminformatics* **2011**, *3*, 33.

502 51. Laggner, C. SMARTS patterns for functional group classification. *Inte: Ligand Software-Entwicklungs und*
503 *Consulting GmbH, Maria Enzersdorf, Austria* **2005**.
504 52. Orengo, C.A.; Michie, A.; Jones, S.; Jones, D.T.; Swindells, M.; Thornton, J.M. CATH–a hierarchic
505 classification of protein domain structures. *Structure* **1997**, *5*, 1093–1109.
506 53. Butina, D. Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: a
507 fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer*
508 *Sciences* **1999**, *39*, 747–750.
509 54. Landrum, G. RDKit: Open-source cheminformatics **2006**.
510 55. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug discovery today* **2006**, *11*, 1046–1053.