**ABSTRACT:** Traditional drug discovery methods focus on optimizing the efficacy of a drug against a single biological target of interest for a specific disease. However, evidence supports the multitarget theory, i.e., drugs work by exerting their therapeutic effects via interaction with multiple biological targets, which have multiple phenotypic effects. Analytics of drug–protein interactions on a large proteomic scale provides insight into disease systems while also allowing for prediction of putative therapeutics against specific indications. We present a Python package for analysis of drug–proteome and drug–disease relationships implementing the Computational Analysis of Novel Drug Opportunities (CANDO) platform. The CANDO package allows for rapid drug similarity assessment, most notably via an in-house interaction scoring protocol where billions of drug–protein interactions are rapidly scored and the similarity of drug-proteome interaction signatures is calculated. The package also implements a variety of benchmarking protocols for shotgun drug discovery and repurposing, i.e., to determine how every known drug is related to every other in the context of the indications/diseases for which they are approved. Drug predictions are generated through consensus scoring of the most similar compounds to drugs known to treat a particular indication. Support for comparing and ranking novel chemical entities, as well as machine learning modules for both benchmarking and putative drug candidate prediction is also available. The CANDO Python package is available on GitHub at https://github.com/ram-compbio/CANDO, through the Conda Python package installer, and at http://compbio.org/software/.

**INTRODUCTION**

Drugs and small molecule compounds exert therapeutic effects via the perturbation of multiple macromolecules, especially proteins. Growing evidence suggests small molecule drugs interact with multiple proteins to enact cellular changes, contrary to the "magic bullet" philosophy often practiced in drug discovery.5−10 Therefore, interpreting the totality of protein interactions for drugs provides greater insight into their therapeutic functions, with the potential for more efficient drug discovery. In addition, drug repurposing has emerged as a valuable alternative to traditional drug discovery pipelines, potentially easing the burden associated with common clinical trial failures.11−14 Multiple groups have taken a multitarget approach for predicting drug effects; both Liu and Altman and Zhou et al. computed interactions between large libraries of drugs and proteins to map targets to side effect outcomes.15,16 Similarly, Simon et al. mapped drug–protein interactions to "effect profiles", of which a given effect is a drug class (for example, calcium channel blocker or stimulant) as opposed to a disease or side effect.17 Numerous groups have used network or systems biology approaches for large scale prediction of drug–disease associations, typically using known drug–protein interactions;18−20 however, no studies have computed proteomic interactions for all drugs for the purpose of benchmarking and prediction for every disease to our knowledge.

We have developed the Computational Analysis of Novel Drug Opportunities (CANDO) platform for analysis of drug interactions on a proteomic scale, adhering to multitarget drug theory,1−7 for the purposes of shotgun drug discovery and repurposing, i.e., to evaluate every drug for every disease. An overview of the platform is provided in Supporting Figure 1. CANDO version 2 (v2) is comprised of a library of 14 606 sequence nonredundant (p-value 10e−7) protein structures extracted from the Protein Data Bank, 2162 human-approved drugs from DrugBank, and 2178 indications/diseases from the Comparative Toxicogenomics Database (CTD), encompassing 18 709 drug–indication associations.21−23 An additional set of 5317 human only protein structures is also available. The platform relates small molecules based on their computed
interactions with all protein structures, known as an interaction signature, then assesses a drug repurposing accuracy based on how similar drug–proteomic signatures are for those drugs approved to treat the same indications. The hypothesis is that drugs with similar interaction signatures will have similar behavior in biological systems and will therefore be useful against the same indications.

Here, we present cando.py, a Python package implementing the CANDO platform for convenient analyses of drug–protein interaction signatures with the ultimate goal of making novel putative drug candidate generation easy and accessible. The package may be used for validation of virtual screening methods for applications in drug discovery and repurposing and for extending or developing novel drug discovery and repurposing platforms. The package reads in a matrix of precomputed interaction scores with any number of proteins, along with a drug to indication mapping, which are then benchmarked. Compound–protein interaction signatures for novel compounds/drugs not present in our library are quickly computed and added to the matrix using our default interaction scoring protocol, allowing for direct comparison and ranking relative to other drug signatures in the platform. The package can also read in any drug–drug similarity/distance matrix computed using any third party package, which may be benchmarked or used for drug–disease association prediction.

**METHODS: CANDO PLATFORM IMPLEMENTATION**

The CANDO platform is implemented in Python as a series of parallel pipelines with modules for the following major protocols (Supporting Figure 1).

**Interaction Scoring Protocol.** The pipelines in the CANDO platform are agnostic to the interaction scoring protocol used: The compound–protein interaction scores in CANDO may be derived from high throughput dissociation constant studies, molecular docking simulations, and/or other
quantification of structure–activity relationships.\textsuperscript{24–26} If more than one protocol is used, then it constitutes a different pipeline within the platform. The reference/default compound-protein interaction scores in the CANDO v2 matrices are computed using a bioinformatic docking protocol that compares the structures of query drugs to all ligands known to bind to a given site on a protein.\textsuperscript{7} Specifically, the COACH algorithm is used to elucidate potential binding sites on each query protein, which uses a consensus approach via three different complementary algorithms that consider substructure or sequence similarity to known binding sites in the PDB.\textsuperscript{27} COACH output includes a set of cocrystallized ligands for each potential binding site, which are then compared to a compound/drug of interest using chemical fingerprinting methods that binarize the presence or absence of particular molecular substructures. The maximum Tanimoto coefficient between the binary vectors of the query compound and the set of all predicted binding site ligands for a protein serve as a proxy for the binding strength. The final output is a series of interaction scores between every drug/compound and every protein structure in the corresponding libraries.

**Benchmarking Protocols.** Each drug/compound is ranked relative to all others based on the pairwise similarity of their proteomic signatures, calculated using the root-mean-square deviation (RMSD) by default, resulting in a ranked list of most similar compounds. By default, all proteins in the library are used for the RMSD calculation but their composition may be varied to allow for more specific queries, both generally or on a per indication basis, which also applies to the canpredict module (discussed below). Other distance metrics, such as cosine distance, may also be used.

The benchmarking protocol (implemented in the canbenchmark module) utilizes a hold-one-out scheme to compute an accuracy for each indication. For a given indication, each approved drug is held-out and the most similar compounds (within various cutoffs) are checked to see if they are also approved for the indication (Figure 2). This protocol is run iteratively and averaged across all indications with two or more drugs approved to provide a drug repurposing accuracy at each cutoff. Both the average indication accuracy (described above) and the pairwise accuracy (the weighted average based on the number of compounds approved for the disease) are outputted, as well as the coverage, which is the number of indications with nonzero accuracy. Benchmarking performance across different versions/pipelines is available in Supporting Figure 2.

**Putative Drug Candidate Generation (Prediction) Protocol.** The ranked lists of most similar compounds to each drug, other than those that are used for benchmarking, are investigated as potential novel treatments. A consensus scoring approach is utilized where for each drug associated with a specific indication, the number of times a particular drug shows up within a certain cutoff of each list is counted. The prediction module canpredict then ranks the top compounds by their consensus scores. Figure 1 provides an example with malaria (Plasmodium falciparum). The top consensus scoring drugs include lumefantrine, a known antimalarial drug, and...
pantoprazole, a proton pump inhibitor that has shown antimalarial activity. Another strong candidate is amino-glutethimide, an aromatase inhibitor with uses including Cushing’s syndrome and various cancers. The exact set of proteins used for the drug—drug similarity calculations can be modified, e.g. specifying only Plasmodium proteins.

**Putative Indication Prediction Protocol.** The canpredict module can also accept a small molecule compound as input, including novel chemical entities, and suggest novel indications for which they may be useful. First the proteomic signature is computed for all proteins in the platform, then the signature is compared to all other drugs in the platform. The most similar drugs to the query compound within a specified cutoff are probed for a consensus among the diseases for which they are indicated, correlative to the disease-focused canpredict module discussed above. Figure 2 presents the results for both an approved drug, ribavirin, and an investigational compound, LMK-235. Ribavirin receives a consensus score of two at the top10 cutoff for both Breast Neoplasms (MeSH:D001943) and Leukemia, Myeloid, Acute (MeSH:D015470), which is supported by clinical trials for both diseases in which ribavirin is the primary intervention.

The three drugs contributing to these consensus scores are gemcitabine, azacitidine, and decitabine, which are all nucleoside analog anticancer therapies. LMK-235 is an investigational histone deacetylase inhibitor that is yet to begin human trials. The canpredict module output with a top20 cutoff includes both Pain (MeSH:D010146) and Hypertension (MeSH:D006973), which are both supported by in vivo experiments.31,32

**AI/Machine Learning Protocols.** The CANDO package also provides support for several machine learning protocols that learn more complex relationships hidden in the drug-proteome interaction signatures to improve performance. The currently supported protocols include support vector machines (SVMs), 1-class SVMs, random forests, and logistic regression, though the latter two are prioritized as they offer insight into feature importance. The modules are trained on the input data to generate models that yield prediction pipelines that are benchmarked using a protocol similar to the one used by canbenchmark: for a given indication, each approved drug is held out while the model is trained on all other drugs. First the proteomic signature is computed for all proteins in the platform, then the signature is compared to all other drugs in the platform. The most similar drugs to the query compound within a specified cutoff are probed for a consensus among the diseases for which they are indicated, correlative to the disease-focused canpredict module discussed above. Figure 2 presents the results for both an approved drug, ribavirin, and an investigational compound, LMK-235. Ribavirin receives a consensus score of two at the top10 cutoff for both Breast Neoplasms (MeSH:D001943) and Leukemia, Myeloid, Acute (MeSH:D015470), which is supported by clinical trials for both diseases in which ribavirin is the primary intervention.

The three drugs contributing to these consensus scores are gemcitabine, azacitidine, and decitabine, which are all nucleoside analog anticancer therapies. LMK-235 is an investigational histone deacetylase inhibitor that is yet to begin human trials. The canpredict module output with a top20 cutoff includes both Pain (MeSH:D010146) and Hypertension (MeSH:D006973), which are both supported by in vivo experiments.31,32

**Development and Implementation.** The CANDO software is available in Python 2.7, 3.6, and 3.7. It is available for installation via the Python Anaconda installer. All data necessary for the benchmarking and prediction modules are available for download directly in the package. The source code, API document, and a Jupyter Notebook tutorial are available on GitHub at https://github.com/ram-compbio/CANDO as well as on http://compbio.org/software/.

**DISCUSSION AND CONCLUSION**

For interaction scoring, in addition to the bioinformatic docking protocol described above, a compound-proteome interaction matrix generated using our state of the art docking program CANDOCK33 with predicted binding energies will be available for use shortly. Indeed, the platform can accept protein–compound interaction, and compound–compound similarity, matrices generated by any method (virtual docking, molecular fingerprinting, gene expression changes, etc.) and benchmark their utility for shotgun drug discovery and repurposing. This is especially useful given that molecular docking and chemical fingerprinting techniques vary greatly in performance.34–36 A Web server hosted on compbio.org that will feature many of the functionalities described is under development.

The multitarget approach to drug discovery is vastly unexplored and shows promise for identifying novel treatments for various diseases based on the results we have obtained using our software. The CANDO Python package allows users to investigate drug–protein interactions on a proteomic scale for the purposes of shotgun drug discovery and repurposing, moving away from the single target and single indication philosophy. The multitarget approach, which in our platform is represented as the synthesis of many virtual screens, is conducive for understanding drug behavior holistically, which will allow for better elucidation of the therapeutic (and adverse) effects these small molecules exert on biological systems. We anticipate that broader use of this platform will inform researchers about potential lead compounds that may be therapeutic for specific indications, leading to accelerated and more efficient drug discovery.

**ASSOCIATED CONTENT**

**Supporting Information**

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00110.

Jupyter Notebook cando.py tutorial (PDF)

Figure S1. Overview of the CANDO drug discovery and repurposing platform. Figure S2. Benchmarking performance across multiple versions and pipelines in the CANDO platform. Figure S3. AI-CANDO platform performance evaluation using the receiver operating characteristic curve (PDF)

API document (PDF)

SMILES strings (TXT)

**AUTHOR INFORMATION**

**Corresponding Author**

Ram Samudrala — Department of Biomedical Informatics,
University at Buffalo, Buffalo, New York 14120, United States;
Email: ram@compbio.org

**Authors**

William Mangione — Department of Biomedical Informatics,
University at Buffalo, Buffalo, New York 14120, United States;
[orcid.org/0000-0003-0582-4247](https://orcid.org/0000-0003-0582-4247)

Zackary Falls — Department of Biomedical Informatics, University at Buffalo, Buffalo, New York 14120, United States
Gaurav Chopra — Department of Chemistry, Purdue Institute for Drug Discovery, Integrated Data Science Institute, Purdue University, West Lafayette, Indiana 47907, United States; orcid.org/0000-0003-0942-7898

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00110

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Liana Bruggemann, Matthew Hudson, Manoj Mammen, and Jim Schuler for their contributions and testing of the CANDO software. This work was supported in part by a National Institute of Health Directors Pioneer Award (DP1OD006779), a National Institute of Health Clinical and Translational Sciences (NCATS) Award (UL1TR001412), NCATS ASPIRE Design Challenge Awards, a National Library of Medicine T15 Award (T15LM012495), a National Cancer Institute/Veterans Affairs Big Data-Scientist Training Enhancement Program Fellowship in Big Data Sciences, and startup funds from the Department of Biomedical Informatics at the University at Buffalo. Startup funds from the Department of Chemistry at Purdue University, Ralph W. and Grace M. Showalter Research Trust award, the Integrative Data Science Initiative award, and the Jim and Dian Robbers Cancer Research Grant for New Investigators award to Gaurav Chopra, as well as additional support in part by a NCATS Clinical and Translational Sciences Award from the Indiana Clinical and Translational Sciences Institute (UL1TR002529), and the Purdue University Center for Cancer Research NIH grant P30CA023168, are also acknowledged. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

■ REFERENCES


