

cando.py: Open Source Software for Predictive Bioanalytics of Large Scale Drug–Protein–Disease Data

William Mangione, Zackary Falls, Gaurav Chopra, and Ram Samudrala*



Cite This: *J. Chem. Inf. Model.* 2020, 60, 4131–4136



Read Online

ACCESS |



Metrics & More

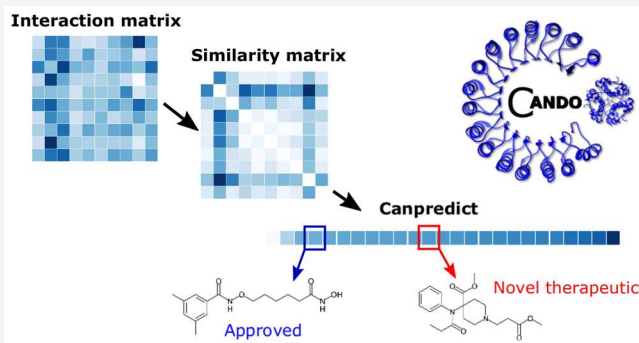


Article Recommendations



Supporting Information

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.^{8–10} 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 proteo-

mic 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

Special Issue: New Trends in Virtual Screening

Received: January 30, 2020

Published: June 9, 2020



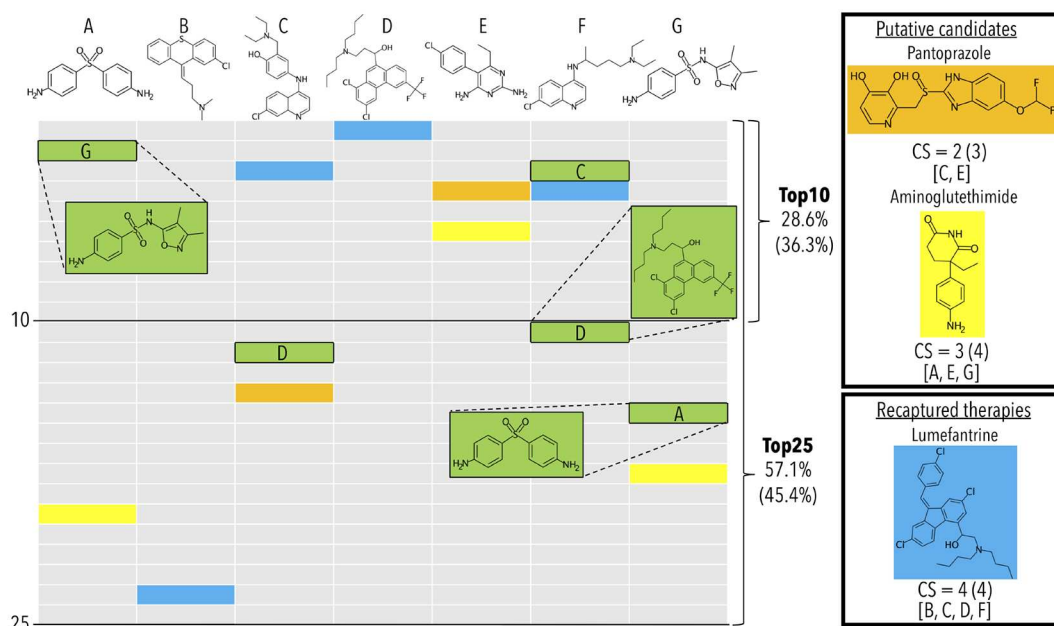


Figure 1. Example of benchmarking and putative therapeutic prediction with the canbenchmark and canpredict modules for malaria (*Plasmodium falciparum*). A subset of the 22 drugs associated with the indication Malaria Falciparum (MeSH:D016778) are labeled A through G, which from left to right are dapsone, chlorprothixene, amodiaquine, halofantrine, pyrimethamine, chloroquine, and sulfisoxazole. The remaining 15 drugs are excluded for illustrative purposes only. The benchmarking accuracies for the top10 and top25 cutoffs, and the top25 consensus scores (CSs) shown in the figure are based on using only the 7 drug subset (top) and for all 22 drugs (bottom, in parentheses). All drugs and table cells in green are used for calculating benchmarking performance, while the yellow, blue, and orange cells are the predictions highlighted by the canpredict module. The columns represent the ranked order of the most similar drugs/compounds based on root-mean-square-deviations of their proteomic signatures to their respective A through G labeled drug above. The drug–proteome interaction matrix used for this example was created from the interaction scores of 5317 human protein structures from the Protein Data Bank with a library of 2162 approved drugs from DrugBank. The benchmarking method tallies the percent of times another drug associated with the indication is captured within a certain column cutoff rank to a held-out compound (A–G) also associated with the indication. In the example, both drugs A and F, dapsone and chloroquine, recapture another drug associated with the indication within the top10 cutoff, which are sulfisoxazole (G) and amodiaquine (C), respectively. This results in a top10 accuracy of 28.6% (two out of seven). Both amodiaquine (C) and sulfisoxazole (G) recapture another associated drug at the top25 cutoff, which are halofantrine (D) and dapsone (A), respectively. This raises the top25 accuracy to 57.1% (four out of seven). This process is iterated over all indications to calculate global accuracies at each cutoff. The canpredict module utilizes a consensus voting scheme to suggest putative drug repurposing candidates based on the similarity of their proteomic interaction signatures to each known treatment for a disease. A tally is kept of how many times a specific drug is captured within a set cutoff to each known treatment (top25 in this example). Pantoprazole (orange), which has shown antimalarial activity in the literature, falls at rank 14 and 4 for amodiaquine (C) and pyrimethamine (E), respectively, receiving a CS of 2. The aromatase inhibitor aminoglutethimide (yellow) has a CS of 3, which we are suggesting as a novel candidate treatment for malaria. Lumefantrine (blue), a known malaria treatment which in this case was not originally included in the Comparative Toxicogenomics Database drug–indication mapping used by the platform, receives a CS of 4. If lumefantrine was originally included as a treatment, the benchmarking scores would increase to 57.1% and 85.7% at the top10 and top25 cutoffs, respectively, which highlights the importance of drug–indication mapping veracity. The benchmarking module provides insight into how well the given drug–protein interaction scoring method is relating drugs in the context of disease, while the canpredict module suggests putative drug repurposing candidates based on drug–drug similarities.

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 com-

pounds/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 disassociation constant studies, molecular docking simulations, and/or other

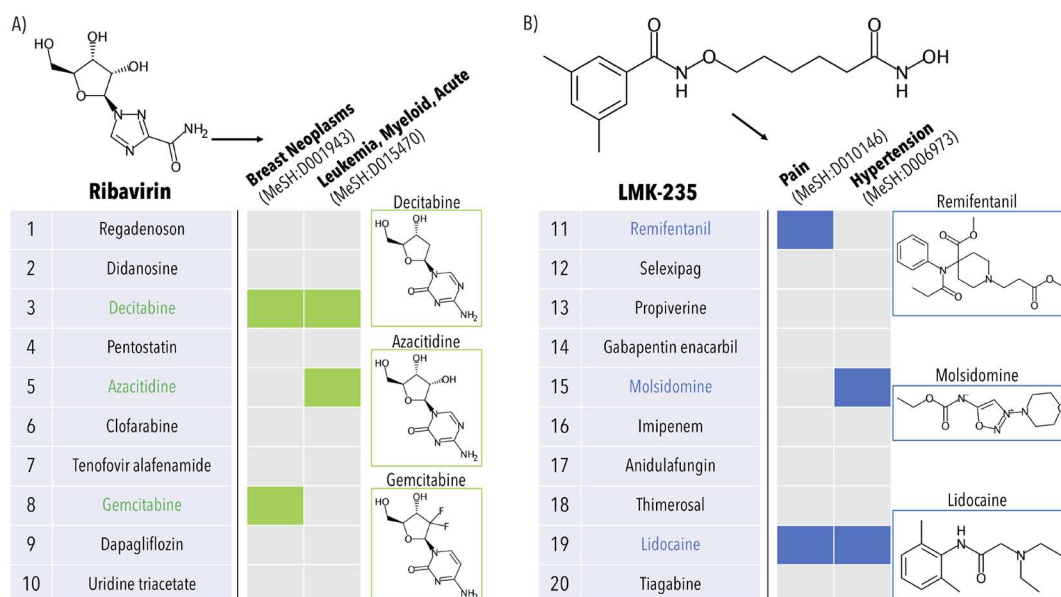


Figure 2. Example of indication prediction using the canpredict module with drugs (ribavirin and LMK-235). The canpredict module also accepts a drug/compound as input and suggests indications for which it may be useful based on the ranked list of the most similar drugs and the diseases for which they are approved. (A) Results for ribavirin, a known antiviral compound, using a set of 5317 human protein structures from the Protein Data Bank to construct the drug–proteome interaction signatures. First, the top10 most similar drugs to ribavirin are computed via the root-mean-square-deviation of their proteomic interaction signatures. Then, the consensus scores of the indications associated with the top10 drugs are calculated. In the example, both Breast Neoplasms (MeSH:D001943) and Leukemia, Myeloid, Acute (MeSH:D015470) receive a consensus score of 2; these two diseases are highlighted as ribavirin is currently in clinical trials for both and has already shown clinical efficacy against acute myeloid leukemia. The three drugs contributing to the consensus scores, decitabine, azacitidine, and gemcitabine, are all chemotherapeutic nucleoside analogs. (B) Results for LMK-235, an experimental histone deacetylase inhibitor currently not approved for human use. In this case, the top20 drugs were probed for a disease consensus (only ranks 11–20 are shown for illustrative purposes). Two indications of note, namely Pain (MeSH:D010146) and Hypertension (MeSH:D006973), both receive a consensus score of 2; there exist multiple studies in the literature supporting both the analgesic and hypotensive properties of LMK-235. The drugs contributing to the consensus scoring include remifentanyl, molsidomine, and lidocaine, as pictured. The CANDO canpredict module is swiftly able to assess the behavioral similarity between the known antiviral ribavirin and several antineoplastic agents, as well as between the experimental compound LMK-235 and analgesic/hypotensive drugs.

quantification of structure–activity relationships.^{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.⁵ 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.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.²⁸ Another strong candidate is aminoglutethimide, 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.^{29,30} 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 approved for the indication in an iterative fashion. In other words, the number of binary classifiers trained corresponds to the number of drugs associated with a particular indication. An equal number of neutral samples are chosen as negative samples during training (except 1-class SVMs), which represent drugs/compounds not associated with the indication. Several metrics are calculated based on the number of times the samples are correctly classified, including the area under the receiver operating characteristic (AUC-ROC) and precision recall (AUPR) curves (Supporting Figure 3). AUC-ROC and AUPR are only available with logistic regression and random forests as they offer classification probabilities. The user may also make predictions for novel or nonassociated compounds after training the classifier on all approved drugs for a particular indication, and both AI/machine learning benchmarking and prediction protocols are amenable to any kind of feature input (e.g., molecular substructures instead of compound–protein interaction scores).

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

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 Diann 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

- (1) 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.
- (2) Sethi, G.; Chopra, G.; Samudrala, R. Multiscale modelling of relationships between protein classes and drug behavior across all diseases using the CANDO platform. *Mini-Rev. Med. Chem.* **2015**, *15*, 705–717.
- (3) Schuler, J.; Samudrala, R. Fingerprinting CANDO: Increased Accuracy with Structure-and Ligand-Based Shotgun Drug Repurposing. *ACS Omega* **2019**, *4*, 17393–17403.
- (4) Mangione, W.; Samudrala, R. Identifying protein features responsible for improved drug repurposing accuracies using the CANDO platform: Implications for drug design. *Molecules* **2019**, *24*, 167.
- (5) Falls, Z.; Mangione, W.; Schuler, J.; Samudrala, R. Exploration of interaction scoring criteria in the CANDO platform. *BMC Res. Notes* **2019**, *12*, 318.
- (6) Fine, J.; Lackner, R.; Samudrala, R.; Chopra, G. Computational chemoproteomics to understand the role of selected psychoactives in treating mental health indications. *Sci. Rep.* **2019**, *9*, 1–15.
- (7) Chopra, G.; Kaushik, S.; Elkin, P. L.; Samudrala, R. Combating ebola with repurposed therapeutics using the CANDO platform. *Molecules* **2016**, *21*, 1537.
- (8) Xie, L.; Evangelidis, T.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. *PLoS Comput. Biol.* **2011**, *7*, e1002037.
- (9) Bolognesi, L. Polypharmacology in a single drug: multitarget drugs. *Curr. Med. Chem.* **2013**, *20*, 1639–1645.
- (10) Horst, J. A.; Laurenzi, A.; Bernard, B.; Samudrala, R. Computational multitarget drug discovery. *Polypharm. Drug Discovery* **2012**, 263.
- (11) Yella, J.; Yaddanapudi, S.; Wang, Y.; Jegga, A. Changing trends in computational drug repositioning. *Pharmaceuticals* **2018**, *11*, 57.
- (12) Pushpakom, S.; Iorio, F.; Eyers, P. A.; Escott, K. J.; Hopper, S.; Wells, A.; Doig, A.; Williams, T.; Latimer, J.; McNamee, C.; Norris, A.; Sanseau, P.; Cavalla, D.; Pirmohamed, M. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discovery* **2019**, *18*, 41.
- (13) Peyvandipour, A.; Saberian, N.; Shafi, A.; Donato, M.; Draghici, S. A novel computational approach for drug repurposing using systems biology. *Bioinformatics* **2018**, *34*, 2817–2825.
- (14) Paranjpe, M. D.; Taubes, A.; Sirota, M. Insights into computational drug repurposing for neurodegenerative disease. *Trends Pharmacol. Sci.* **2019**, *40*, 565.
- (15) Liu, T.; Altman, R. B. Relating essential proteins to drug side-effects using canonical component analysis: a structure-based approach. *J. Chem. Inf. Model.* **2015**, *55*, 1483–1494.
- (16) Zhou, H.; Gao, M.; Skolnick, J. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Sci. Rep.* **2015**, *5*, 11090.
- (17) Simon, Z.; Peragovics, G.; Vigh-Smeller, M.; Csukly, G.; Tombor, L.; Yang, Z.; Zahoránszky-Köhalmi, G.; Végner, L.; Jelinek, B.; Hári, P.; et al. Drug effect prediction by polypharmacology-based interaction profiling. *J. Chem. Inf. Model.* **2012**, *52*, 134–145.
- (18) Zeng, X.; Zhu, S.; Liu, X.; Zhou, Y.; Nussinov, R.; Cheng, F. deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* **2019**, *35*, 5191–5198.
- (19) Saberian, N.; Peyvandipour, A.; Donato, M.; Ansari, S.; Draghici, S. A new computational drug repurposing method using established disease-drug pair knowledge. *Bioinformatics* **2019**, *35*, 3672–3678.
- (20) Qabaja, A.; Alshalalfa, M.; Alanazi, E.; Alhaji, R. Prediction of novel drug indications using network driven biological data prioritization and integration. *J. Cheminf.* **2014**, *6*, 1.
- (21) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.
- (22) Davis, A. P.; Grondin, C. J.; Johnson, R. J.; Sciaky, D.; King, B. L.; McMorran, R.; Wieggers, J.; Wieggers, T. C.; Mattingly, C. J. The comparative toxicogenomics database: update 2017. *Nucleic Acids Res.* **2017**, *45*, D972–D978.
- (23) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *International Tables for Crystallography Vol. F: Crystallography of biological macromolecules*; Springer, 2006; pp 675–684.
- (24) Vaidya, A.; Jain, S.; Jain, S.; Jain, A. K.; Agrawal, R. K. Quantitative structure-activity relationships: a novel approach of drug design and discovery. *J. Pharm. Sci. Pharmacol.* **2014**, *1*, 219–232.
- (25) Hall, M. D.; Yasgar, A.; Peryea, T.; Braisted, J. C.; Jadhav, A.; Simeonov, A.; Coussens, N. P. Fluorescence polarization assays in high-throughput screening and drug discovery: a review. *Methods Appl. Fluoresc.* **2016**, *4*, 022001.
- (26) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for molecular docking: a review. *Biophys. Rev.* **2017**, *9*, 91–102.
- (27) Yang, J.; Roy, A.; Zhang, Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, *29*, 2588–2595.
- (28) Riel, M. A.; Kyle, D. E.; Bhattacharjee, A. K.; Milhous, W. K. Efficacy of proton pump inhibitor drugs against *Plasmodium falciparum* in vitro and their probable pharmacophores. *Antimicrob. Agents Chemother.* **2002**, *46*, 2627–2632.
- (29) Assouline, S.; Culjkovic, B.; Cocolakis, E.; Rousseau, C.; Beslu, N.; Amri, A.; Caplan, S.; Leber, B.; Roy, D.-C.; Miller, W. H.; Borden, K. L. B. Molecular targeting of the oncogene *lF4E* in acute myeloid leukemia (AML): a proof-of-principle clinical trial with ribavirin. *Blood* **2009**, *114*, 257–260.

(30) Pettersson, F.; Yau, C.; Dobocan, M. C.; Culjkovic-Kraljacic, B.; Retrouvay, H.; Puckett, R.; Flores, L. M.; Krop, I. E.; Rousseau, C.; Cocolakis, E.; et al. Ribavirin treatment effects on breast cancers overexpressing eIF4E, a biomarker with prognostic specificity for luminal B-type breast cancer. *Clin. Cancer Res.* **2011**, *17*, 2874–2884.

(31) Lin, T.-B.; Hsieh, M.-C.; Lai, C.-Y.; Cheng, J.-K.; Chau, Y.-P.; Ruan, T.; Chen, G.-D.; Peng, H.-Y. Modulation of Nerve Injury-induced HDAC4 Cytoplasmic Retention Contributes to Neuropathic Pain in Rats. *Anesthesiology* **2015**, *123*, 199–212.

(32) Choi, S. Y.; Kee, H. J.; Sun, S.; Seok, Y. M.; Ryu, Y.; Kim, G. R.; Kee, S.-J.; Pflieger, M.; Kurz, T.; Kassack, M. U.; Jeong, M. H. Histone deacetylase inhibitor LMK235 attenuates vascular constriction and aortic remodelling in hypertension. *J. Cell. Mol. Med.* **2019**, *23*, 2801–2812.

(33) Fine, J.; Konc, J.; Samudrala, R.; Chopra, G. CANDOCK: Chemical atomic network based hierarchical flexible docking algorithm using generalized statistical potentials. *J. Chem. Inf. Model.* **2020**, *60*, 1509–1527.

(34) Li, J.; Fu, A.; Zhang, L. An overview of scoring functions used for protein-ligand interactions in molecular docking. *Interdiscip. Sci.: Comput. Life Sci.* **2019**, *11*, 320.

(35) Saikia, S.; Bordoloi, M. Molecular docking: challenges, advances and its use in drug discovery perspective. *Curr. Drug Targets* **2019**, *20*, 501–521.

(36) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry: miniperspective. *J. Med. Chem.* **2014**, *57*, 3186–3204.

cando.py: Open source software for predictive bioanalytics of large scale drug-protein-disease data

William Mangione¹⁺, Zackary Falls¹, Gaurav Chopra², Ram Samudrala^{1*}

*Corresponding author

RS email: ram@compbio.org

¹Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, 14120, United States

²Department of Chemistry, Purdue Institute for Drug Discovery, Integrated Data Science Institute, Purdue
University, West Lafayette, IN, 47907, United States

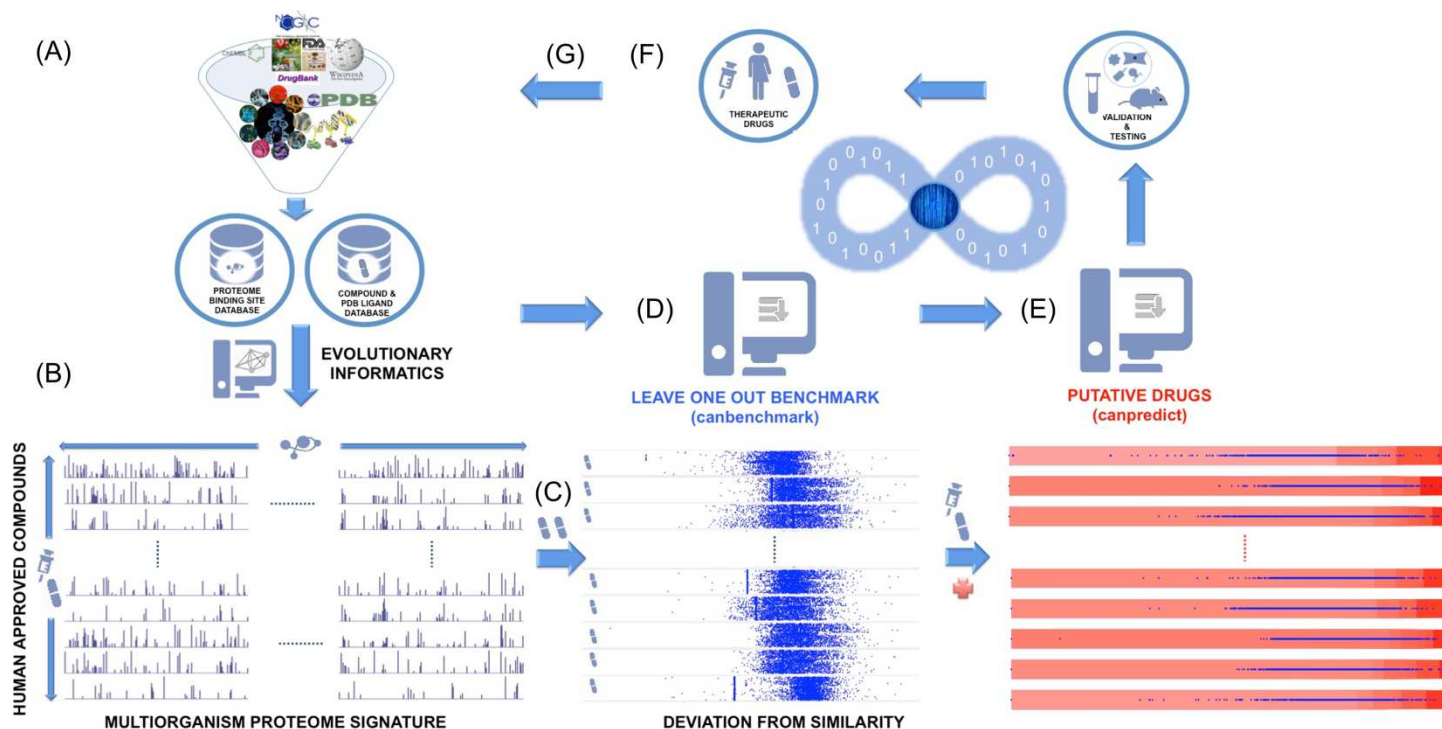


Figure S1: Overview of the CANDO drug discovery and repurposing platform. (A) Data collection: drugs, small molecule compounds, and protein structures are collected from public databases, most notably DrugBank and the Protein Data Bank (PDB). Different protein sets are collated, including the multiorganism sequence non-redundant PDB set (BLAST p-value cutoff $10e-7$) of 14,606 chains, a human-only set of 5,317 chains, and various pathogenic proteomes. (B) Interaction scoring protocol: interactions are computed between all compounds and all proteins using our bioinformatic docking method or virtual docking simulations, which essentially results in a virtual interaction screen for every protein chain against every drug/compound in the library. The bioinformatic docking protocol generates billions of drug-protein interaction scores rapidly (see Methods). (C) Drug comparison protocol: drug-drug similarity is assessed via computing the similarity (typically by calculating the root-mean-square deviation (RMSD)) between two drug-proteome interaction signatures, which allows each drug to be ranked relative to each other (based on the composition of their interactions). The exact protein set to be considered for the RMSD calculation can be modified in various ways. (D) Benchmarking protocol: the platform is benchmarked using known drug-indication associations as a gold standard, primarily from the Comparative Toxicogenomics Database. Briefly, an accuracy is calculated based on the number of times another drug approved to treat a disease is captured within a certain rank in the similarity list of a hold-out drug known to treat the same disease. This is repeated for all indications in the platform with at least two drugs associated and averaged at various cutoffs. (E) Putative drug candidate generation protocol: putative drug candidates for a specific indication or disease are predicted based on the similarity of their proteomic interaction signatures to those of drugs known to treat that indication, and ranked via a consensus voting scheme where the top compounds are prioritized if they are highly ranked in multiple similarity lists of approved drugs for that indication. (F) Validation pipeline: strong candidates proceed to validation studies, including in vitro and in vivo experiments, with the ultimate goal of conducting clinical trials for FDA approval. (G) Platform optimization: All results from the validation studies are fed back into the platform, using machine learning to optimize performance.

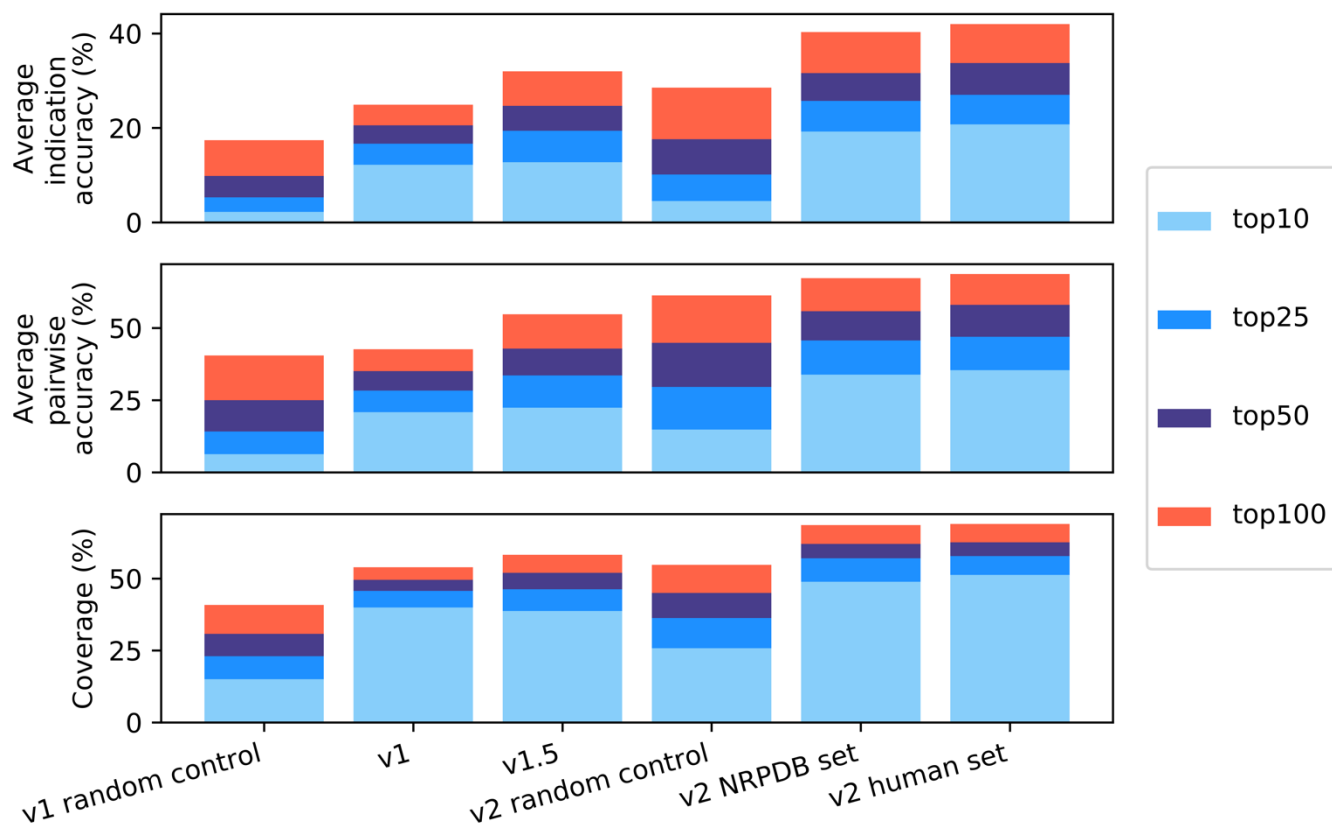


Figure S2: Benchmarking performance across multiple versions and pipelines in the CANDO platform. The standard three metrics assessed by the benchmarking protocol across various versions/pipelines are pictured, including average indication accuracy (top), average pairwise accuracy (middle), and coverage (bottom). Average indication accuracy is the average of each individual indication accuracy, the average pairwise accuracy is the weighted average of each indication accuracy based on the number of compounds/drugs associated with the indications, and coverage is the percent of indications with non-zero accuracy scores. The v1 and v1.5 pipelines differ only in the drug-protein interaction scoring protocol (see Falls et al. 2019), though all data including drugs, indications, and proteins is consistent. The total number of indications with greater than 2 drugs associated in v1 and v2 are 1439 and 1570, respectively. The number of drugs in v1 and v2 are 3,733 and 2,162, respectively. The number of proteins in v1 is 46,784; v2 features two protein sets: the multiorganism sequence nonredundant set of 14,606 and a set of 5,317 human protein structures, both from the Protein Data Bank. The v1 random control is computed via randomizing values in a 3,733x46,784 matrix, computing the root-mean-square-deviation between all randomized vectors, and performing the benchmarking analysis (see Methods). The result reported above is the average of 100 iterations. The v2 random control pictured was generated via a process similar to v1, but with an initial matrix of 2,162x14,606. The colors indicate using a cutoff of 10 (light blue), 25 (blue), 50 (purple), or 100 (red) for the benchmarking analysis. The results of v1 pipelines and v2 pipelines cannot be compared directly to each other as the number of drugs and drug-disease associations changes between versions. The v2 human protein set pipeline currently performs the best in terms of v2, achieving the best scores in all three metrics. The benchmarking analysis helps to indicate which pipelines are relating drugs approved for the diseases more accurately, which will ultimately help to generate novel drug repurposing candidates more efficiently.

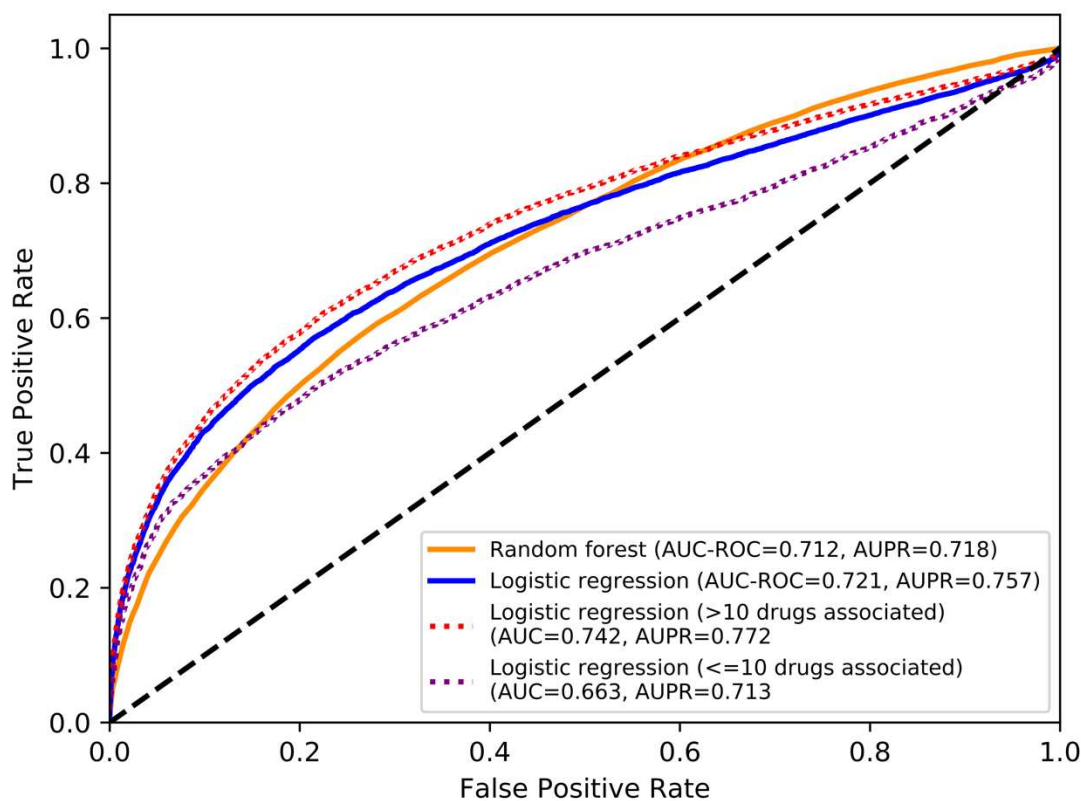


Figure S3: AI-CANDO platform performance evaluation using the receiver operating characteristic (ROC). Shown are the results using logistic regression and random forest binary classifiers across 1,570 indications. A leave-one-out cross validation scheme is used where each drug approved for a disease is held-out and tested on a model trained with every other drug approved for that disease (i.e. 18,101 models are trained corresponding to the number of drug-disease associations for diseases with two or more drugs associated). An equal number of negative samples are chosen by randomly selecting from the set of drugs not associated with the disease. In each case, the positive class probability is outputted and the threshold is varied to generate the receiver operating characteristic curve (ROC), area under the curve (AUC-ROC), and area under the precision-recall curve (AUPR). The dotted black line indicates the expected ROC curve if guessing randomly, which corresponds to AUC-ROC and AUPR values of 0.5. As shown, the logistic regression models (orange) outperform the random forest models (blue) in both AUC-ROC and AUPR. The AI-CANDO pipelines are based on machine learning models trained using the set of 5,317 human protein structures from the Protein Data Bank to construct the drug-proteome interaction signatures. The dotted lines indicate the difference in performance when considering indications with more than ten approved drugs (red) or with a maximum of ten approved drugs (purple), indicating the importance of sample sizes for drug-indication prediction. The model training is amenable to any input features outside of drug-protein interactions, such as molecular substructures or gene expression signatures. Though the models perform better than random, continued development of both input features (enhanced drug-protein interaction scoring) and model architecture will further improve the AI/machine learning benchmarking and prediction module.