Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

Article

Multiscale virtual screening optimization for shotgun drug repurposing using the CANDO platform

Matthew L. Hudson [1] and Ram Samudrala [1,*]

[1] Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences at the University at Buffalo, State University of New York, Buffalo, New York 14203, United States of America

* Correspondence: ram@compbio.org; Tel.: +1-716-888-4858

Abstract: Drug repurposing, the practice of utilizing existing drugs for novel clinical indications, has tremendous potential for improving human health outcomes and increasing therapeutic development efficiency. The goal of multidisease multitarget drug repurposing, also known as shotgun drug repurposing, is to develop platforms that assess the therapeutic potential of each existing drug for every clinical indication. Our Computational Analysis of Novel Drug Opportunities (CANDO) platform for shotgun multitarget repurposing implements several pipelines via large scale modelling and simulation of interactions between comprehensive libraries of drugs/compounds and protein structures. In these pipelines, each drug is described by an interaction signature that is then compared to all other signatures that are then sorted and ranked based on similarity. Pipelines within the platform are benchmarked based on their ability to recover known drugs for all indications in our library, and predictions are generated based on the hypothesis that (novel) drugs with similar signatures may be repurposed for the same indication(s). The drug-protein interactions in the platform used to create the drug-proteome signatures may be determined by any screening or docking method but the primary approach used thus far has been an in house similarity docking protocol. In this study, we calculated drug-proteome interaction signatures using the publicly available molecular docking method Autodock Vina and created hybrid decision tree pipelines that combined our original bio- and cheminformatic approach with the goal of assessing and benchmarking their drug repurposing capabilities and performance. The hybrid decision tree pipeline outperformed the corresponding two docking-based pipelines it was synthesized from, yielding an average indication accuracy of 13.3% at the top10 cutoff (the most stringent), relative to 10.9% and 7.1% for its constituent pipelines, and a random control accuracy of 2.2%. We demonstrate that docking based virtual screening pipelines have unique performance characteristics and that the CANDO shotgun repurposing paradigm is not dependent on a specific docking method. Our results also provide further evidence that multiple CANDO pipelines can be synthesized to enhance drug repurposing predictive capability relative to their constituent pipelines. Overall, this study indicates that pipelines consisting of varied docking based signature generation methods can capture unique and useful signal for accurate comparison of drug-proteome interaction signatures, leading to improvements in the benchmarking and predictive performance of the CANDO shotgun drug repurposing platform.

Keywords: drug repurposing; virtual screening; multiscale; multitargeting; polypharmacology; computational biology; drug repositioning; structural bioinformatics; molecular docking; proteomic signature

1. Introduction

Drug repurposing

Pharmacological innovation reduces human mortality rates and provides substantial improvements to quality of life[1]. Therapeutic compounds that have been discovered, lab tested preclinically, and evaluated for risks and efficacy in clinical trials are approved by regulatory bodies

© 2021 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

such as the United States FDA for specific indications[2]. Potential failures impose high opportunity costs, and realities of market forces and investment distort the types of ailments for which treatments are pursued[3–5]. The rate of novel drug discovery has been slowing as costs have been increasing, illustrating the need for more efficient paradigms [6].

Drug discovery traditionally relies on screening a compound or set of compounds against a biological target, typically a protein for a specific indication. Generally, these approaches incorporate high throughput *in vitro* compound screens[7] and/or cell-based assays[8] of candidates drawn from wet laboratory studies or computational screens of virtual representations of compounds and biological targets[9–11]. If promising *in vitro* leads are found, they undergo *in vivo* testing, eventually leading to approval for clinical use if they continue to demonstrate relative efficacy and safety[2]. Traditional drug discovery methods tend to be focused on a single target and indication[11,12]. However, drugs and other human ingested compounds interact promiscuously with many proteins in the body[13,14]. These off target interactions are responsible for side effects and the fact that one drug may be useful for treating multiple indications[15–19]. Single target approaches may miss promising leads and potentially beneficial off-target side effects, while drugs that have already been discovered and vetted for safety may be used in novel treatment contexts.

Drug repurposing is the practice of finding new uses for existing drugs, taking advantage of prior safety, efficacy, and pharmacological knowledge and data[15]. Drug repurposing has the potential to arbitrarily increase the utility of the FDA approved drug library[17,20,21], particularly via innovations such as multitarget drug repurposing[22,23]. Drug repurposing has yielded new uses for multiple drugs[15,17] and has demonstrated potential for the treatment of viral[24,25], bacterial[26], and complex indications such as cancer[17].

Computational drug repurposing using molecular docking

Computational models that improve drug discovery and repurposing leverage rapidly increasing computer processing power and vast collections of preclinical (*in vitro*, *in vivo*) and clinical data[22]. Although there are a variety of computational approaches, the most relevant ones to this study are structure-based. Structure-based approaches focus on modelling/simulating the effects the three-dimensional (3D) structure of a compound may have on one or more macromolecules, typically protein structures[27]. Structure representations are based on data obtained from x-ray diffraction, NMR spectroscopy, cryogenic electron microscopy, and biochemical and biophysical simulation studies. These models may incorporate other features such as predicted protein-compound binding sites, simulations of the surrounding chemical environments, and functional characteristics of protein structures.

Molecular docking models the three-dimensional (3D) interaction between small molecule compounds and macromolecular protein structures[28–32]. Typically, these simulations algorithmically calculate the optimal position and orientation of a compound structure that interacts with (or binds to) a particular region of a protein structure, and its corresponding interaction strength, using physics-based[33] or knowledge-based[34,35] forcefields or scoring functions. The characteristics of a correctly modelled compound-protein structure provides researchers insight into the biological implications of the interaction: for example, a researcher may infer that a signalling pathway may be interrupted if a particular protein were to be inhibited by the compound based on the strength of its binding energy[36]. Molecular docking is also useful when researching large sets of compounds and proteins[37]. By comparing the relative differences in interactions between protein-compound pairs, the researcher can rank and organize pairs according to the strength of their interaction score and/or their similarity to identify patterns that are apparent only when examining large sets with many possible combinations, which is difficult and expensive to do *in vitro* or *in vivo* experiments[22,38]. Molecular docking techniques have varying performance advantages and limitations[39,40]; however, provided that docking approaches are used wisely in concert with other experimental techniques, they have the potential to be useful for drug repurposing, particularly in a large scale context[38].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

Shotgun multitarget multidisease drug repurposing using the CANDO platform

The Computational Analysis of Novel Drug Opportunities (CANDO) platform was developed to mitigate endemic problems in drug discovery and and enable multitarget approaches to drug repurposing[22,38,41–45]. The CANDO platform is designed to provide insights about the holistic behavior of compounds interacting within complex biological systems, including how a compound behaves relative to other compounds, and is an extensible standardized framework for building and combining drug repurposing, discovery and design simulation pipelines. Similarity of drug-protein interaction behavior between a small molecule drug/compound and its macromolecular environment is hypothesized to indicate similarity of drug therapeutic function[22]. In traditional structure-based and ligand-based drug discovery, therapeutic similarity inferences are often based on molecular target similarity and compound similarity[46]. CANDO extends the similarity assumption principle to include holistic multiscale interaction similarity, that is, characterizing compounds by the nature of their interaction with entire proteomes and (eventually) interactomes[22,38]. Extending the interaction similarity frontier enables CANDO to account for the promiscuous nature of compound interaction within biological systems and characterize previously unconsidered therapeutic functions of existing approved drugs. CANDO pipelines are evaluated by a benchmarking protocol which examines the relative ranking of every drug for every indication with two or more approved drugs. Analyzing the relative ranking of approved drugs for each indication enables evaluation of the effectiveness of the platform for recovering known information, comparing relative pipeline performance for particular indications, calculating accuracy and precision for ranking approved drugs; and determining which components of the platform need improvement.

In this study we set out to extend the CANDO platform with an additional molecular docking pipeline using the popular software AutoDock Vina[30] to determine whether the prior CANDO performance was dependent on a specific molecular docking protocol, how different molecular docking protocols affect CANDO performance, and whether hybridizing molecular docking pipelines yields improved performance, as we have previously observed combining structure- and ligand-based CANDO pipelines[43].

2. Results

Benchmarking performance of the different pipelines

The performance of two new primary pipelines and a hybrid one in the CANDO platform were investigated, and compared to those previously created, including a random control (see METHODS). The first is the Vina pipeline that used the eponymous molecular docking program to screen the CANDO v1.5 3733 drug/compound library against a 134 protein subset of the full proteome library (Vina-134). Multiple binding sites for each protein were predicted and targeted for docking, and the strongest interaction scores were used to construct the drug-proteome signatures.

The second pipeline used is the default CANDO v1.5 pipeline restricted to the same 134 protein subset (v1.5-134). We generated a hybrid decision tree pipeline drawn from a combination of the Vina-134 and the v1.5-134 pipelines. For comparison, we examined the performance of these pipelines with respect to a random control and the v1.5 pipeline implemented with the full CANDO proteome library consisting of 46,784 protein structures (v1.5-full).

Figure 2 illustrates the relative performance of these different pipelines. At the top10 threshold, the hybrid decision tree yields 13.3% accuracy, v1.5-134 yields 10.9%, and Vina-134 yields 7.11%. The v1.5-134 and v1.5-full pipelines outperform the Vina pipeline but the latter was able to substantially contribute to the superior performance of the hybrid pipeline. Notably, the hybrid decision tree pipeline outperforms the v1.5-full pipeline with a top10 accuracy of 12.8% with two orders of magnitude difference in the number of proteins used in the implementation of the pipeline (134 vs 46,784).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

Divergence in indication accuracy at various thresholds

Figure 3 illustrates the similarity and divergence of indication accuracy performance at various thresholds: i.e. instances where the Vina-134 pipeline outperforms the v1.5-134 pipeline, instances where the v1.5-134 pipeline outperforms the Vina-134 pipeline, instances where each pipeline yields the same indication accuracy, and instances where each pipeline yields zero percent accuracy. At the top10 threshold, the Vina-134 pipeline had 191 indications (about 13% of all indications) that outperformed the v1.5 pipeline, which had 363 indications outperform Vina-134 (about 25% of all indications). There were 885 equivalently performing indications (with 828 of them at zero percent accuracy) at the top10 cutoff. Overall, the divergence in relative performance increased as the thresholds became less stringent (the CANDO pipeline outperformance share began to decline slightly after the top 5% threshold). v1.5 had a larger share of indications in which it outperformed Vina-134. After the top50 cutoff, the proportion of equivalent indication accuracies that are both zero relative to the total equivalent indication accuracies begins to decline rapidly.

Net differences in indication accuracy

Figure 4 elucidates the net differences in top10 accuracies between two pipelines (and the proportion of approved drugs recovered per indication in the top10) for 700 indications. With some notable exceptions, the v1.5-134 pipeline outperforms the Vina-134 pipeline in frequency and magnitude. On a per indication basis, as the total number of approved drugs decreases, the Vina-134 pipeline starts to have a higher share of the outperforming indications in terms of frequency and magnitude.

Relative pipeline indication accuracy

Pipelines differ at average indication accuracy thresholds, and on a per indication basis. In some cases, a pipeline that performs worse overall can may do better for a specific indication. Figure 3, Figure 4, and Figure 5 illustrate the overall divergence, the magnitude of divergence, and the threshold frequency distributions. On a per indication basis there is divergence in the relative indication accuracy at various cutoffs, both in terms of net difference in accuracy, recovery at a particular threshold, and frequency of a particular indication being recovered at a particular interval. The divergence between the per indication performance of each pipeline elucidated by Figures 2, 3, and 4 suggests that each pipeline should be used in conjunction with one another for maximum indication inclusivity and accuracy.

Comparison of pipeline distribution of per indication accuracies

Figure 5 illustrates the distribution of indication accuracies by counting the frequency each indication falls within a certain accuracy range. The dissimilarity of pipeline distributions at each cutoff was assessed by applying the Kolmogorov-Smirnoff test. The v1.5 pipeline outperforms Vina-134 overall (which yields a higher frequency of indications exceeding 50% accuracy).

Indication accuracy distribution

Figure 6 examines the distribution of indications to illustrate their relative performance within each pipeline. Pipelines can also be compared with symmetrical accuracy distribution charts, where individual pipeline accuracy is denoted along the x and y axes. Each point can represent a particular indication (e.g. one of the 1439 indications in the CANDO platform), a defined indication class (e.g. all 39 indications with the string “neoplasm”), or some other way of denoting indications (e.g. indications that occupy a particular branch of the Medical Subjects Heading (MeSH) classification [47] or those that are ontologically similar [48]). When pipelines reach accuracy consensus (or near consensus) for a particular indication (or indication grouping), the point falls on or close to the 45 degree symmetry line. These figures suggest that different pipelines have varying success in benchmarking performance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

on a per-indication basis. More rigorous clustering analysis, indication classification, and indication definition will yield deeper insight into the relative strengths of each pipeline.

Distribution of individual drug-indication pair rankings

Supplementary figure S1 plots every drug-indication pair and its corresponding rank within each pipeline. These suggest that there is some substantial ranking consensus between each pipeline as well as substantial divergence. The distribution was plotted at linear and logarithmic scale to illustrate the density of approved drug-indication pair ranking consensus and divergence. There is a high density of drug-indication pairs that have relatively high ranking in each pipeline. There is also a high density of drug-indication pairs that have a significantly higher ranking in the v1.5-134 pipeline than the Vina-134 pipeline. As with pipeline per-indication accuracy divergence, further investigation into drug-indication pair divergence may help improve performance of individual and hybrid pipelines, particularly in cases where one pipeline ranked a drug-indication pair substantially higher than the other one (e.g. top100 in one and bottom 50% in the other).

3. Discussion

Multiple large scale virtual screening pipelines

In this investigation, we hypothesised that distinct docking methods would yield distinct drug-proteome interaction signatures due to differing simulation implementation, and correspondingly differing performance for shotgun drug repurposing: The default similarity docking in CANDO is a knowledge-based template/comparative modelling protocol [22,38,41,44,45] and AutoDock is a more traditional molecular docking approach with physics based force fields [30]. Including other molecular docking pipelines beyond the default pipeline implemented in the platform enables us to evaluate whether or not CANDO as a platform was specifically dependent on the drug-proteome signature generation methodology implemented in the default pipeline.

Our results demonstrate that CANDO platform is not dependent on a single pipeline implementation, and also that combining different virtual screening pipelines can yield better performance relative to using the individual ones. On a platform level, the drug-proteome signature ranking and indication recovery paradigm is viable using more than one means of signature generation. On a pipeline level, the pipelines (the two large scale virtual screening pipelines and the combined decision tree pipeline) each demonstrate a varying degrees of performance and instances of unique signal capture.

The Vina-134 pipeline implemented in this study was viable in that it performed substantially better than the random control and performed at a significant fraction of the performance of the original default pipeline that utilized a much larger protein set. However, small protein libraries have been shown to perform relatively well and some subsets of protein libraries perform better than others [44]. As previously demonstrated[43], hybrid pipelines can draw from the strengths of each constituent pipeline. As is the case here, the absolute performance of the Vina-134 pipeline was not the best, yet it substantially contributed to the higher performing hybrid pipeline.

Limitations and future work

Although some pipelines yield superior signal over others in specific circumstances, precisely identifying why this occurs warrants further investigation. On a per indication basis, it is possible to identify superior performance of one pipeline over another (Figures 3 and 4), but the MeSH indication classes are not precisely defined or have varying levels of specificity to one another. This issue will be addressed in the future through the use of more precisely defined indication mapping, for instance by using a realism based ontology[48,49]. We are also using mathematical, statistical and machine learning techniques to rigorously evaluate and enhance CANDO pipeline performance, as well as to identify clusters of drug-indication pair rankings when comparing different pipelines and methods

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

[43,50], to yield insight into the ability of each pipeline to accurately recover known per indication association information and make useful predictions for downstream prospective preclinical and clinical validation [42,51].

4. Materials and Methods

CANDO platform and pipeline implementation

Figure 1 provides an overview of the CANDO platform and the particular pipeline implementations relevant to this study. The platform uses drug/compound and protein structure libraries curated from public sources, and implements protocols for drug- and compound-proteome interaction signature generation, signature similarity calculation and sorting, assessing whether known drugs are ranked highly for the correct indications for single or hybrid pipelines (benchmarking) and generating novel putative drug candidates for specific indications (prediction). CANDO relative drug ranking pipelines have utility in many drug repurposing research contexts. For example, these pipelines can be used for lead generation for subsequent *in vitro*/*in vivo* testing and eventual off-label clinical use by physicians. By assessing the top ranking subset of drugs, a researcher or clinician can efficiently infer promising experimental or clinical drug candidates based on relative drug ranking to FDA approved drug treatments and prior experimental evidence. For many clinical indications, CANDO pipelines are able to identify and highly rank FDA approved drug treatments along with drugs that are FDA approved for other indications. Researchers can also infer associations between clinical indication classes, diseases, and biological pathways through examination of indication-indication association networks connected by highly ranked drugs they have in common or other features of their respective compound-proteome signature. As illustrative examples of the broad uses of CANDO, Supplementary Figure S2 and Supplementary Table S1 describe the indication-indication associations for a selection of MeSH neoplasm indications based on shared drugs ranked in the Top 10 in the Vina pipeline.

Drug/compound, protein structure, and indication library curation

The default CANDO pipelines are implemented using bio/cheminformatic docking protocols, where interactions are predicted from curated drug and protein libraries. The specific implementations and evolution of the libraries has been reported previously extensively in several publications [22,38, 41,44,45]. Briefly, the initial versions of CANDO (v1 and v1.5) incorporated 46,784 proteins and 2030 indication associations for 1439 drugs (out of 3733 compounds total). Much of the data was drawn from the Protein Data Bank[52], Food and Drug Administration, PubChem[53], the Comparative Toxicogenomics Database[54], DrugBank[55], protein structure modeling[56], and other sources.

The pipelines used in this study rely on curated sublibraries of structures of 3733 drugs/compounds and 134 proteins, and 13,746 drug-indication mappings, obtained from the same sources as above. We used the sublibraries to rapidly evaluate the utility of multiple molecular docking pipelines.

Drug- and compound-proteome interaction signature generation

A CANDO virtual screening pipeline simulates the interactions between all of its proteins and drugs/compounds, usually 3D structures, and is not dependent on any particular approach to accomplish this. These simulations generate proteomic similarity signatures (the vector of drug-protein interaction scores). The default CANDO platform pipelines generate drug-proteome interaction signatures using bioinformatic and cheminformatic docking protocols also described elsewhere extensively [22,38,41,44,45]. These signatures are compared for similarity and ranked. CANDO pipeline version 1.5 [45] is a refinement of the original default pipeline [22,38,41,44] that used near identical libraries but improved interaction scoring [45]. We extended the drug- and compound-proteome interaction signature protocols to include the calculated binding energies

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

generated by the program AutoDock Vina [30] as well as created hybrid pipelines combining molecular docking with similarity docking (further details below).

Drug- and compound-proteome signature similarity calculation and sorting

Broadly, the CANDO platform works by sorting every drug/compound relative to every other one based on their similarity and then uses known drug-indication associations to assess performance (Figure 2). Various pipelines implemented in CANDO generate drug-proteome interaction signatures for similarity sorting [22,38,43–45]. Underlying this platform is the core assumption that similarity of drug interaction behavior across a proteome may be used to infer similarity in therapeutic function. The similarity between each drug and every other drug/compound is calculated using the root mean square deviation of the individual interaction scores across a pair of drug-proteome interaction signatures [38].

Combined drug-proteome interaction signatures form an interaction matrix, with drugs along one axis and proteins on the other. These signatures are compared with one another and then ranked on a per drug basis, and the quality of the resulting ranking evaluated using the leave-one-out benchmarking protocol described below.

Benchmarking CANDO platform pipelines

Our benchmarking protocol calculates performance for every indication with at least two approved drugs (1439 out of 3733 total) at various cutoffs, considering only the the top 10 (abbreviated as "top10"), top25, top37 or 1%, top50, top100, top5%, top10%, and top50% of similarly ranked drugs. For each indication, the accuracy is derived from calculating how many known drugs mapped to that indication were "recovered" and highly ranked at various cutoffs.

We utilize three metrics to benchmark pipeline performance: average indication accuracy, pairwise accuracy, and coverage[22,38,41,44,45], all assessed at the different cutoffs. Average (mean) indication accuracy (%) is the average of all individual indication accuracies. The individual indication accuracy metric is calculated using the formula $c/d \times 100$, where c is a count of the number of times at least one approved drug for the indication was recovered within a particular cutoff and d is the total number of drugs approved for that indication. The other two benchmarking metrics are pairwise accuracy (weighted average of indication accuracies using the total number of approved drugs per indication) and coverage (number of indications that have an accuracy greater than zero).

New and hybrid pipelines

The pipelines examined in this study are derived from similarity ranking and benchmarking drug-proteome interaction signatures generated by large scale bioinformatic and molecular docking. The CANDO platform is not limited to using docking based virtual screening pipelines, and has the potential to incorporate many different approaches to pipeline implementation and data sets (for example, ligand centric approaches have proven quite effective [43]).

Virtual screening pipeline using AutoDock Vina

We used a small sublibrary (134 proteins) of the full CANDO proteome library to create the new molecular docking virtual screening pipeline due to computational constraints, and also because we previously have shown that appropriately selected sublibraries of similar size from the full library yields similar or better benchmarking performance [44]. We used the popular software AutoDock Vina version 1.1.2 [30] for molecular docking of each protein structure against 3733 drugs/compounds from the CANDO v1.5 libraries. As with the similarity docking, we used COFACTOR[57] to predict binding sites, for binding search space size optimization[58], and used the strongest interaction score (lowest calculated binding energy) for each simulation from multiple sites. The best interaction score values for a drug-protein pair were used to generate the drug-proteome signatures.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

Decision tree pipeline

Prior CANDO platform investigations have demonstrated that multiple pipelines can be combined into a hybrid decision tree to maximize indication accuracy by drawing from pipelines that produce the best performance on a per indication basis [43]. We used a similar approach in this investigation, using the pipeline that had the highest performance at the top10 cutoff.

Controls

We also compared benchmarking performance of the pipelines to values obtained using a hypergeometric distribution that estimates the numerical probability of making a correct prediction by chance. This is one of the random control reference benchmarks used in the CANDO platform, the implementation of which is covered in detail in prior publications [43,45]. Benchmarking performance was also compared to the default pipeline implementations in CANDO version 1 and version 1.5 using the complete libraries.

5. Conclusions

Our results indicate that the utilization of multiple diverse docking based virtual screening approaches in drug repurposing contexts such as the CANDO platform improves benchmarking performance. The Vina-134 pipeline performance indicates that the CANDO platform hypothesis of drug behavior similarity is not limited to the original similarity docking protocol for interaction signature generation. The hybrid decision tree pipeline performance provides further evidence that multiple signature generation pipelines may be combined to yield improved performance. Ongoing and future platform enhancement will incorporate multiple signature generation protocols and pipeline synthesis using AI/machine learning approaches to optimize performance. These improvements in turn will lead to greater predictive power and higher confidence in novel drug candidates generated for specific indications, which will be verified via prospective preclinical and clinical studies.

Supplementary Materials: The following are available online at http://www.mdpi.com//xx/1/5/s1 and http://compbio.buffalo.edu/data/mc_cando_multiscale_optimization/, Supplementary figure S1: Plot of every drug-indication pair and its corresponding rank within each pipeline. Supplementary figure S2: Indication-indication associations between MeSH Neoplasm associated classes based on the number of compounds predicted in the Top 10 cutoff by the Vina pipeline. Supplementary table S3: Raw indication-indication association counts.

Funding: This work was supported in part by a National Institutes of Health Director's Pioneer Award 279 (DP1OD006779), a National Institutes of Health Clinical and Translational Sciences Award (UL1TR001412), a NCATS ASPIRE Design Challenge Award, and startup funds from the Department of Biomedical Informatics at the University at Buffalo.

Acknowledgments: Additional support provided by the Center for Computational Research at the University at Buffalo. The authors thank James Schuler, Will Mangione, Dr. Zackary Falls, Liana Bruggemann, and Dr. Manoj Mammen for their valuable input during the development of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CANDO	Computational Analysis of Novel Drug Opportunities
FDA	Food and Drug Administration
NMR	Nuclear Magnetic Resonance

References

1. Lichtenberg, F.R. Pharmaceutical innovation, mortality reduction, and economic growth. Technical report, National Bureau of Economic Research, 1998.
2. for Drug Evaluation, C.; Research. Drug Development Approval Process.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

3. Mullard, A. New drugs cost US[Dollars] 2.6 billion to develop, 2014.
4. DiMasi, J.A.; Hansen, R.W.; Grabowski, H.G. The price of innovation: new estimates of drug development costs. *Journal of health economics* **2003**, *22*, 151–185.
5. DiMasi, J.A. New drug development in the United States from 1963 to 1999. *Clinical Pharmacology & Therapeutics* **2001**, *69*, 286–296.
6. Scannell, J.W.; Blanckley, A.; Boldon, H.; Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature reviews Drug discovery* **2012**, *11*, 191.
7. Broach, J.R.; Thorner, J.; others. High-throughput screening for drug discovery. *Nature* **1996**, *384*, 14–16.
8. Michelini, E.; Cevenini, L.; Mezzanotte, L.; Coppa, A.; Roda, A. Cell-based assays: fuelling drug discovery. *Analytical and bioanalytical chemistry* **2010**, *398*, 227–238.
9. 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.
10. Lionta, E.; Spyrou, G.; K Vassilatis, D.; Cournia, Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current topics in medicinal chemistry* **2014**, *14*, 1923–1938.
11. Patel, C.N.; Georrge, J.J.; Modi, K.M.; Narechania, M.B.; Patel, D.P.; Gonzalez, F.J.; Pandya, H.A. Pharmacophore-based virtual screening of catechol-o-methyltransferase (COMT) inhibitors to combat Alzheimer's disease. *Journal of Biomolecular Structure and Dynamics* **2018**, *36*, 3938–3957.
12. Medina-Franco, J.L.; Giulianotti, M.A.; Welmaker, G.S.; Houghten, R.A. Shifting from the single to the multitarget paradigm in drug discovery. *Drug discovery today* **2013**, *18*, 495–501.
13. L Bolognesi, M. Polypharmacology in a single drug: multitarget drugs. *Current medicinal chemistry* **2013**, *20*, 1639–1645.
14. Hu, Y.; Bajorath, J. Monitoring drug promiscuity over time. *F1000Research* **2014**, *3*.
15. Ashburn, T.T.; Thor, K.B. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery* **2004**, *3*, 673–683.
16. Langedijk, J.; Mantel-Teeuwisse, A.K.; Slijkerman, D.S.; Schutjens, M.H.D. Drug repositioning and repurposing: terminology and definitions in literature. *Drug discovery today* **2015**, *20*, 1027–1034.
17. Palumbo, A.; Facon, T.; Sonneveld, P.; Blade, J.; Offidani, M.; Gay, F.; Moreau, P.; Waage, A.; Spencer, A.; Ludwig, H.; others. Thalidomide for treatment of multiple myeloma: 10 years later. *Blood, The Journal of the American Society of Hematology* **2008**, *111*, 3968–3977.
18. Roth, B.L.; Sheffler, D.J.; Kroeze, W.K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature reviews Drug discovery* **2004**, *3*, 353–359.
19. de Lera, A.R.; Ganesan, A. Epigenetic polypharmacology: from combination therapy to multitargeted drugs. *Clinical epigenetics* **2016**, *8*, 105.
20. Arts, E.J.; Hazuda, D.J. HIV-1 antiretroviral drug therapy. *Cold Spring Harbor perspectives in medicine* **2012**, *2*, a007161.
21. Sardana, D.; Zhu, C.; Zhang, M.; Gudivada, R.C.; Yang, L.; Jegga, A.G. Drug repositioning for orphan diseases. *Briefings in bioinformatics* **2011**, *12*, 346–356.
22. 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.
23. Mangione, W.; Falls, Z.; Chopra, G.; Samudrala, R. cando. py: Open source software for analyzing large scale drug-protein-disease data. *bioRxiv* **2019**, p. 845545.
24. Xu, M.; Lee, E.M.; Wen, Z.; Cheng, Y.; Huang, W.K.; Qian, X.; Julia, T.; Kouznetsova, J.; Ogden, S.C.; Hammack, C.; others. Identification of small-molecule inhibitors of Zika virus infection and induced neural cell death via a drug repurposing screen. *Nature medicine* **2016**, *22*, 1101.
25. Schuler, J.; Hudson, M.L.; Schwartz, D.; Samudrala, R. A systematic review of computational drug discovery, development, and repurposing for Ebola virus disease treatment. *Molecules* **2017**, *22*, 1777.
26. Roder, C.; Thomson, M.J. Auranofin: repurposing an old drug for a golden new age. *Drugs in R&D* **2015**, *15*, 13–20.
27. Ou-Yang, S.s.; Lu, J.y.; Kong, X.q.; Liang, Z.j.; Luo, C.; Jiang, H. Computational drug discovery. *Acta Pharmacologica Sinica* **2012**, *33*, 1131–1140.
28. Taylor, R.D.; Jewsbury, P.J.; Essex, J.W. A review of protein-small molecule docking methods. *Journal of computer-aided molecular design* **2002**, *16*, 151–166.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

29. Pagadala, N.S.; Syed, K.; Tuszynski, J. Software for molecular docking: a review. *Biophysical reviews* **2017**, *9*, 91–102.
30. Trott, O.; Olson, A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **2010**, *31*, 455–461.
31. Fine, J.; Konc, J.; Samudrala, R.; Chopra, G. CANDOCK: Chemical atomic network based hierarchical flexible docking algorithm using generalized statistical potentials. *BioRxiv* **2019**, p. 442897.
32. Yuriev, E.; Ramsland, P.A. Latest developments in molecular docking: 2010–2011 in review. *Journal of Molecular Recognition* **2013**, *26*, 215–239.
33. Huang, N.; Jacobson, M.P. Physics-based methods for studying protein-ligand interactions. *Current Opinion in Drug Discovery and Development* **2007**, *10*, 325.
34. Gohlke, H.; Klebe, G. Statistical potentials and scoring functions applied to protein–ligand binding. *Current opinion in structural biology* **2001**, *11*, 231–235.
35. Muegge, I. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspectives in Drug Discovery and Design* **2000**, *20*, 99–114.
36. Lokhande, K.B.; Nagar, S.; Swamy, K.V. Molecular interaction studies of Deguelin and its derivatives with Cyclin D1 and Cyclin E in cancer cell signaling pathway: The computational approach. *Scientific reports* **2019**, *9*, 1–13.
37. Ma, D.L.; Chan, D.S.H.; Leung, C.H. Molecular docking for virtual screening of natural product databases. *Chemical science* **2011**, *2*, 1656–1665.
38. 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.
39. Rodrigues, J.; Melquiond, A.; Karaca, E.; Trellet, M.; Van Dijk, M.; Van Zundert, G.; Schmitz, C.; De Vries, S.; Bordogna, A.; Bonati, L.; others. Defining the limits of homology modeling in information-driven protein docking. *Proteins: Structure, Function, and Bioinformatics* **2013**, *81*, 2119–2128.
40. Yuriev, E.; Agostino, M.; Ramsland, P.A. Challenges and advances in computational docking: 2009 in review. *Journal of Molecular Recognition* **2011**, *24*, 149–164.
41. Chopra, G.; Samudrala, R. Exploring polypharmacology in drug discovery and repurposing using the CANDO platform. *Current pharmaceutical design* **2016**, *22*, 3109–3123.
42. Chopra, G.; Kaushik, S.; Elkin, P.L.; Samudrala, R. Combating ebola with repurposed therapeutics using the CANDO platform. *Molecules* **2016**, *21*, 1537.
43. Schuler, J.; Samudrala, R. Fingerprinting CANDO: Increased Accuracy with Structure-and Ligand-Based Shotgun Drug Repurposing. *ACS omega* **2019**, *4*, 17393–17403.
44. 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.
45. Falls, Z.; Mangione, W.; Schuler, J.; Samudrala, R. Exploration of interaction scoring criteria in the CANDO platform. *BMC research notes* **2019**, *12*, 318.
46. Cavasotto, C.N.; Phatak, S.S. Homology modeling in drug discovery: current trends and applications. *Drug discovery today* **2009**, *14*, 676–683.
47. Lipscomb, C.E. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* **2000**, *88*, 265.
48. Arp, R.; Smith, B.; Spear, A.D. *Building ontologies with basic formal ontology*; Mit Press, 2015.
49. Schuler, J.; Mangione, W.; Samudrala, R.; Ceusters, W. Foundations for a Realism-based Drug Repurposing Ontology. *10th Annual International Conference on Biomedical Ontology* **2019**.
50. Schuler, J.; Falls, Z.; Mangione, W.; Hudson, M.; Bruggemann, L.; Samdurala, R. Evaluating performance of drug repurposing technologies. *Drug Discovery Today* **2020**, *invited*.
51. Mangione, W.; Falls, Z.; Melendy, T.; Chopra, G.; Samudrala, R. Shotgun drug repurposing biotechnology to tackle epidemics and pandemics. *Drug Discovery Today* **2020**, *Epub ahead of print*. doi:10.1016/j.drudis.2020.05.002.
52. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic acids research* **2000**, *28*, 235–242.
53. Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; others. PubChem substance and compound databases. *Nucleic acids research* **2016**, *44*, D1202–D1213.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

54. 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.; others. The comparative toxicogenomics database: update 2013. *Nucleic acids research* **2012**, *41*, D1104–D1114.
55. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; others. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research* **2010**, *39*, D1035–D1041.
56. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics* **2008**, *9*, 40.
57. Roy, A.; Yang, J.; Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research* **2012**, *40*, W471–W477.
58. Feinstein, W.P.; Brylinski, M. Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets. *Journal of cheminformatics* **2015**, *7*, 1–10.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

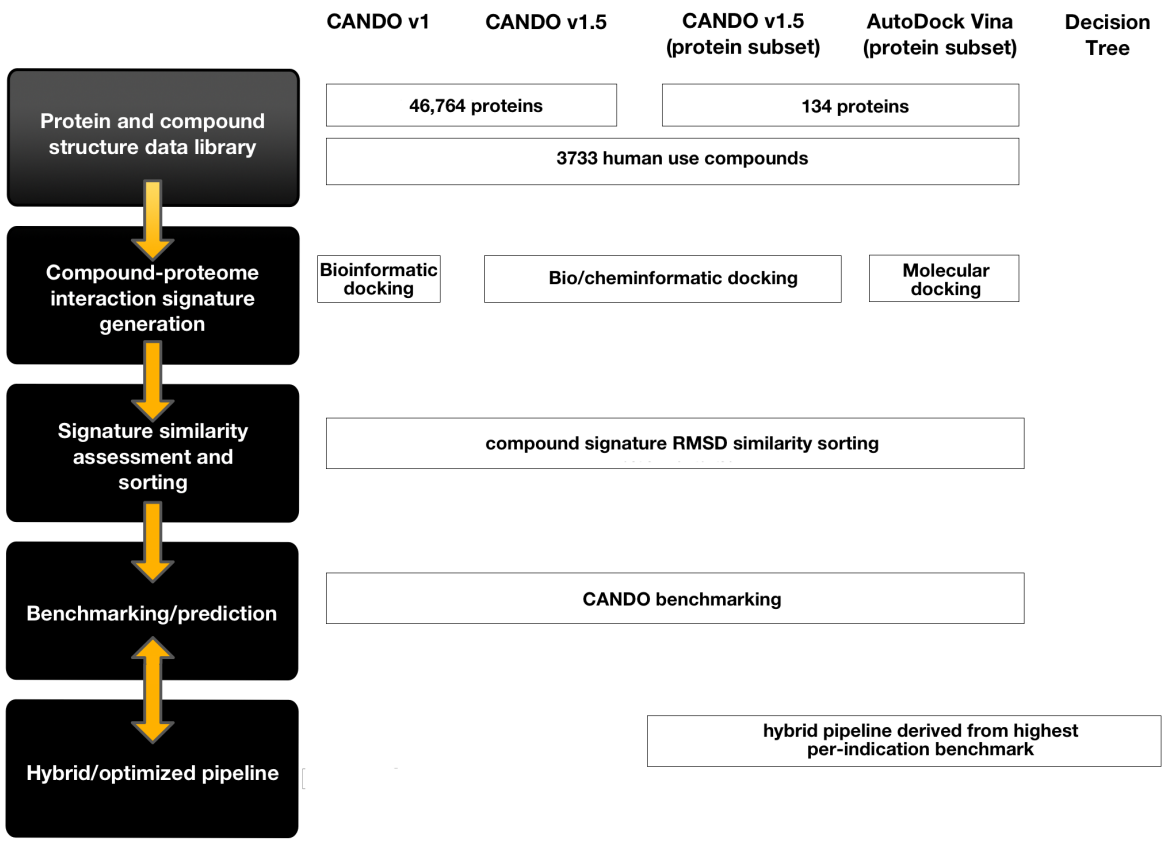

Figure 1. CANDO shotgun drug repurposing platform and pipeline overview. On the left side of the figure is a flow diagram which indicates the general protocol for implementing a CANDO drug-proteome pipeline. To the right of the flow diagram, each pipeline relevant to this investigation is displayed along with implementation details for each phase of the CANDO protocol. To the right of the flow diagram, each pipeline relevant to this investigation is displayed along with implementation details for each phase of the CANDO protocol. (A) **Data curation:** The drug-proteome pipelines utilize libraries of protein structure and drug structure representations. (B) **Interaction scoring protocol:** These pipelines use bioinformatic, cheminformatic, and molecular docking methods to predict the interaction scores of each of the protein and drug representation pair. The set of protein interaction scores for each drug is considered its interaction signature. Each interaction signature can be compared with one another by assessing the root mean square deviation of their interaction signatures. (C) **Drug comparison protocol:** Every drug signature is compared with every other drug signature. After every comparison is made, each drug has a list containing the ordered set of every other drug, from most similar signature to least similar signature. (D) **Benchmarking protocol:** The CANDO benchmarking procedure assesses, for every drug, how many other drugs with the same indication association are found within certain ranking cutoffs. An indication specific accuracy score is produced by averaging the recovery rate of co-associated drugs for every drug associated with the particular indication for particular ranking cutoffs. Overall pipeline average indication accuracy is the mean of all per indication accuracies for a particular cutoff. Three pipelines were generated during this investigation: v1.5 (implemented with a subset of the CANDO proteome library), AutoDock Vina (using the same proteome sublibrary), and a hybrid decision tree pipeline derived from the former two pipelines. Each of the subset pipelines utilized a small sublibrary (134 proteins) of the original CANDO v1 and v1.5 pipelines. Although the pipelines used different signature generation approaches (similarity docking and AutoDock Vina molecular docking), their signatures underwent the same similarity assessment and benchmarking protocol. Although drug-proteome pipelines generally follow the same protocol, there is room for variation such as alternate docking, similarity assessment, and benchmarking approaches.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

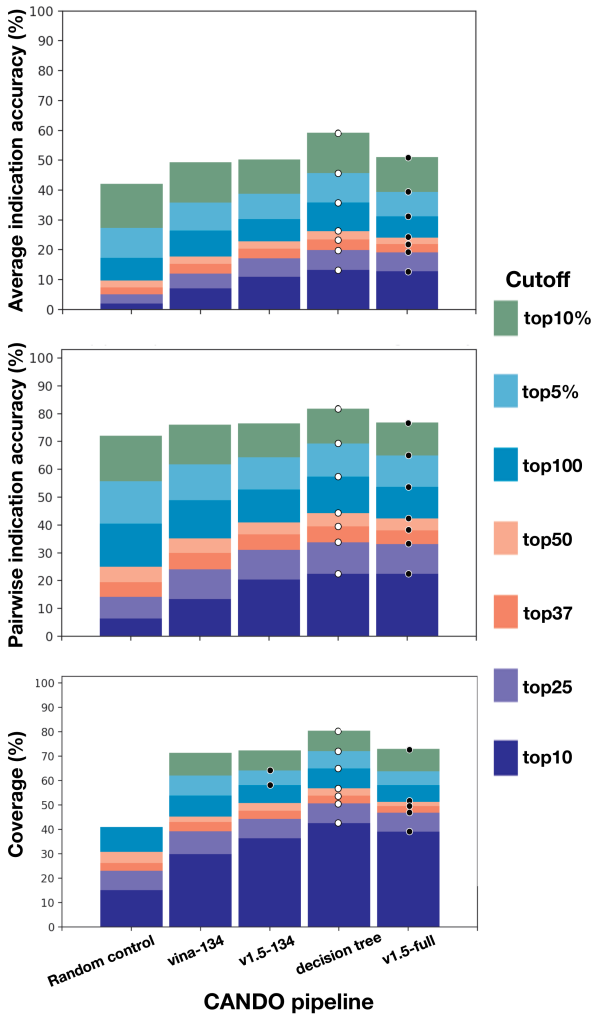

Figure 2. Benchmarking performance of CANDO pipelines used in this study. Three docking derived pipelines implemented in CANDO: v1.5-full, using the interaction scores generated by the default CANDO v1.5 similarity docking protocol for the full (46,784) proteome library; v1.5-134, using the same interaction scoring protocol for a 134 protein sublibrary; and Vina-134, based on interaction scores generated using AutoDock Vina for the 134 protein sublibrary are compared with (d) a hybrid decision tree pipeline derived from combining pipelines (b) and (c) as well as (a) random control reference pipeline calculated numerically from a hypergeometric distribution[43]. The pipelines are assessed by three CANDO platform benchmarking metrics: average indication accuracy (%), pairwise accuracy (%), and coverage (%). Performance cutoffs are denoted by colored bars from most to least stringent: top10 (dark purple), top25 (light purple), top37/top1% (dark pink), top50 (light)pink, top100 (dark blue), top5% (light blue), top10% (dark green), top50% (light green) for 1439 indications with at least two approved drugs using a leave-one-out benchmarking protocol (see METHODS). White dots denote highest overall accuracy at each threshold. The hybrid decision tree pipeline, that incorporates the highest indication accuracies from the Vina-134 and v1.5-134 pipelines performed the best at all cutoffs (white dots). Black dots denote high performance in individual pipelines, which was obtained using the two v1.5 pipelines, one based on the 134 proteome sublibrary and the other on the full proteome library. v1.5-134 yields the highest top50 percent average indication accuracy (85.3%), top 100 coverage (52.742%), top 5% coverage (64.382%), top 50% coverage (96.064%). Individually, the Vina-134 pipeline significantly outperforms the random control and yields a significant fraction of the performance of the v1.5 pipelines. The hybrid decision tree pipeline performed the best, indicating that diversity in pipeline simulation implementation can be leveraged to increase drug repurposing performance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

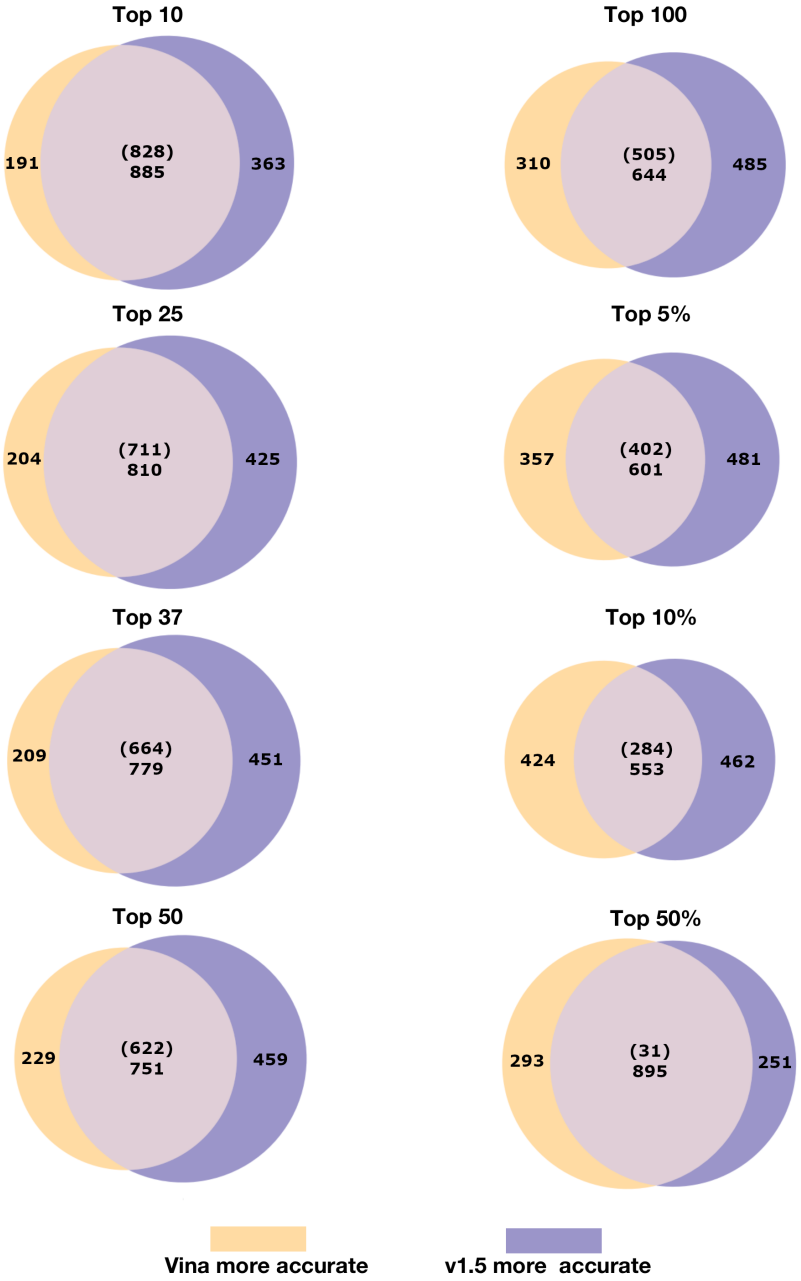

Figure 3. Comparison and contrasting indication accuracies for two CANDO platform pipelines at different cutoffs. Each Venn diagram represents the set of indication accuracies (1439 total) for the Vina-134 and v1.5-134 pipelines at different cutoffs (top10, top25, top37, top50, top100, top5%, top10%, and top50%). Indications that scored higher for the Vina pipeline are in yellow, indications that scored higher for the v1.5 pipeline are in purple, and indications that scored the same for each pipeline are in gray. The number of indications where both pipelines yield 0% accuracy is provided in parentheses and the total number of equal indication accuracies provided below. At the top10 cutoff, the Vina-134 pipeline yields higher accuracies for 191 indications, the v1.5 pipeline for 363 indications, and each pipeline yields the same accuracy for 885 indications. Although the Vina-134 pipeline produces a substantial number of indication accuracies that are higher or equivalent to v1.5 indication accuracies, the v1.5 pipeline outperforms the Vina-134 pipeline at every cutoff except for top50%. Nonetheless, the orthogonality in the above diagrams indicates that individual pipelines can be synergistically combined into a hybrid pipeline that yields considerable performance improvements.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

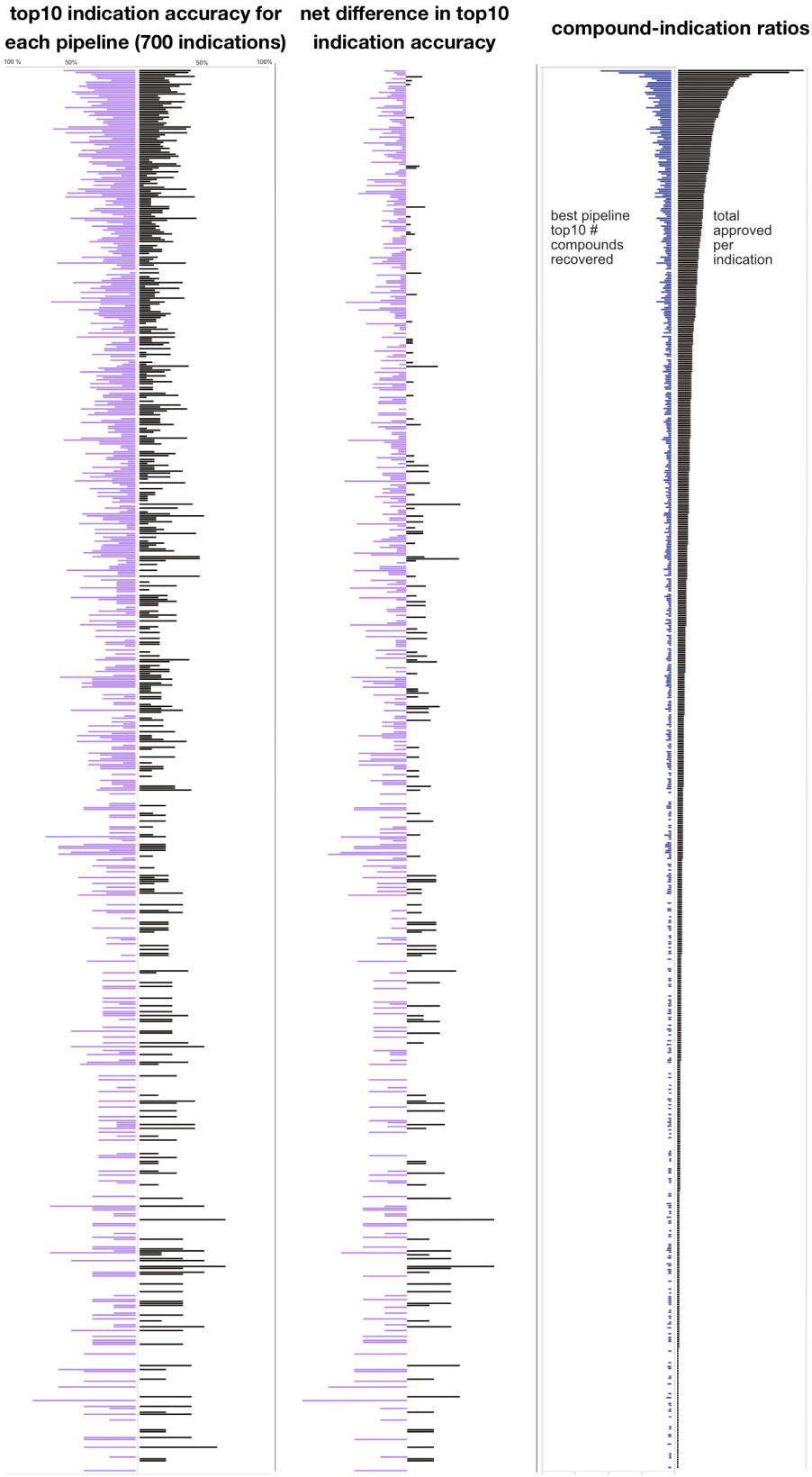

Figure 4. Comparison of 700 indication accuracies for two CANDO platform pipelines at the top10 cutoff. The top10 indication accuracies for 700 indications produced by the Vina-134 and v1.5-134 pipelines are shown in the left panel, with the the v1.5-134 pipeline per indication accuracies in purple on the left side and the Vina-134 pipeline accuracies in black on the right. The net difference in pipeline accuracy for the same indication is shown in the center panel, using the same percentage scale as the left. The number of drugs recovered by the best performing pipeline (in blue on the left side) and the total number of drugs approved per indication (in black on the right side) is shown in the right panel. The number of approved drugs for all three panels ranges from 158 drugs at the top to two drugs at the bottom. Generally, the v1.5-134 pipeline outperforms the Vina-134 pipeline, both by number of indications and net difference in accuracy per indication.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

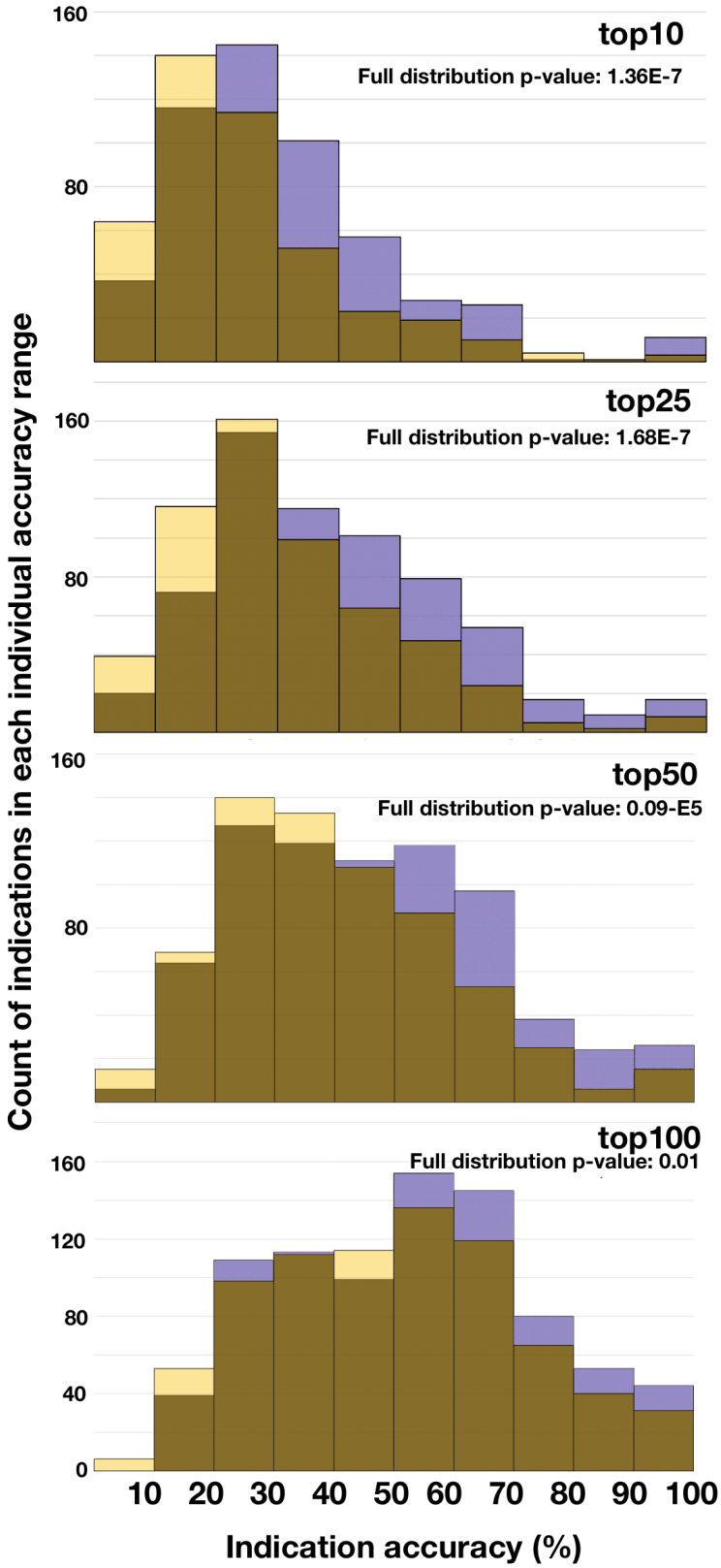

Figure 5. Frequency comparison of indication accuracies for two CANDO platform pipelines at different cutoffs. Shown are four histograms denoting the frequency with which indications fall within particular accuracy ranges (Vina-134 pipeline accuracies are in light yellow, and v1.5 accuracies in black). The similarity of each distribution is assessed by the p-value using the Kolmogorov-Smirnoff test (p-values less than 0.05 are considered to be significant). The v1.5 pipeline outperforms Vina-134 overall but the p-values indicate that the accuracy distributions are different for the two pipelines, indicating the utility of combining pipelines to produce synergistic performance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 April 2021 doi:10.20944/preprints202104.0475.v1

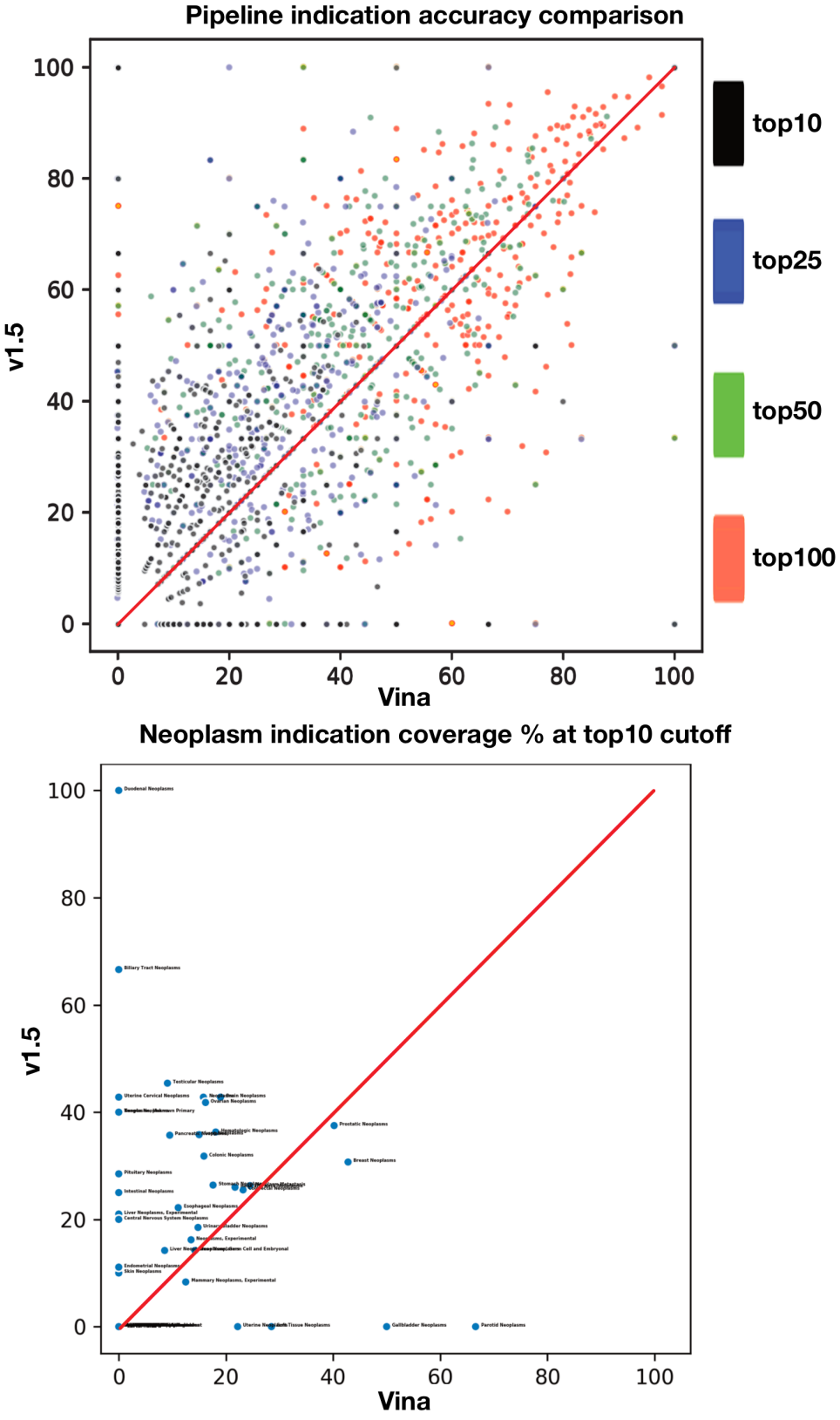

Figure 6. Comparison of indication accuracies at various cutoffs and for a defined indication class for two CANDO platform pipelines. The top panel denotes a symmetrical accuracy chart. Each axis measures the indication accuracy for each pipeline and indications are plotted according to their corresponding accuracies (different cutoffs are distributed in alternate colors). Points that land on the 45 degree red line are indications where the pipelines reached consensus, points that fall closer to a particular axis achieved relatively higher score with the corresponding pipeline. The bottom panel isolates indications for the defined class "Neoplasm" comprising of 39 indications with the corresponding string. The asymmetrical distribution of the accuracy plot suggests pipeline accuracy differentiation, i.e., different pipelines have differing performance strengths and weaknesses, on a per indication and indication class level.