Evaluating the performance of drug-repurposing technologies

James Schuler *, **Zackary Falls, William Mangione, Matthew L. Hudson, Liana Bruggemann, Ram Samudrala** *

Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, USA

Drug-repurposing technologies are growing in number and maturing. However, comparisons to each other and to reality are hindered because of a lack of consensus with respect to performance evaluation. Such comparability is necessary to determine scientific merit and to ensure that only meaningful predictions from repurposing technologies carry through to further validation and eventual patient use. Here, we review and compare performance evaluation measures for these technologies using version 2 of our shotgun repurposing Computational Analysis of Novel Drug Opportunities (CANDO) platform to illustrate their benefits, drawbacks, and limitations. Understanding and using different performance evaluation metrics ensures robust cross-platform comparability, enabling us to continue to strive toward optimal repurposing by decreasing the time and cost of drug discovery and development.

Keywords:
Drug repurposing; Computational drug repurposing;
Drug repositioning; Computational drug repositioning;
Biomedical Informatics; Performance evaluation; Performance metrics;
Area under the curve; Normalized discounted cumulative gain; Validation

James Schuler is an MD-PhD student in the Department of Biomedical Informatics at the Jacobs School of Medicine and Biomedical Sciences at the University at Buffalo. His thesis brings rigor to the field of computational drug repurposing via two goals: examining how performance evaluation in repurposing platforms is biased and developing best practice standards and metrics to overcome these biases to achieve the necessary end goals of determining which drugs work best in relevant contexts. His additional research/clinical interests include ontological realism and pediatric endocrinology, especially diabetes technology and treatment.

Ram Samudrala is a professor and Chief, Division of Bioinformatics, University at Buffalo (2014-present) researching multiscale modeling of protein and proteome structure, function, interaction, design, and evolution and its application to medicine. Previously, Samudrala was faculty at the University of Washington (2001–2014), and postdoctoral fellow at Stanford University (1997–2000).

* *Corresponding authors.* Schuler, J. (jcschule@buffalo.edu), Samudrala, R. (ram@compbio.org).

Introduction

Drug-repurposing technologies allow us to predict new uses for previously approved drugs.[1] Although the drug discovery process normally takes years of work and costs billions of dollars, drug repurposing lowers barriers to the entry of a drug to the market.[2–4] The ultimate goal of repurposing research is to decrease the time and cost of drug discovery and development by accurately predicting clinical utility and using the predictions to improve health and quality of life. Successful instances of drug repurposing have been based on anecdotal evidence, *in vitro* and *in vivo* screening, and discovery of serendipitous positive effects in clinical trials or analysis of patient health records post-market.[5] Drug-repurposing technologies aim to make this process systematic and skip intermediate steps (Fig. 1).

Specific goals of drug repurposing differ based on their eventual utility. For a pharmaceutical company, a specific goal might be to find a single blockbuster drug that changes the default treatment of a condition shared by millions, such as hypertension.[6] A basic science drug-repurposing approach at an academic institution might focus more on public benefit and less on monetary outcomes, such as to find a treatment for an orphan or rare

disease.[7] Less-defined end goals (high risk) enable researchers to systematically disrupt the entire drug discovery and development process (high reward). Drug-repurposing technologies, such as our CANDO platform,[8–18] are used to systematically predict the relative efficacy of every drug in its comprehensive library to treat every disease/indication, minimizing risks, and amplifying rewards. In conjunction with mechanistic basic science analyses, these platforms may be used to better understand the science of drug behavior and thereby model reality with greater fidelity.

Distinctions between noncomputational and computational drug-repurposing technologies are blurring. Experiments designated as computational might rely on data collected in a bench environment, such as protein–ligand binding energy measurement and gene expression studies.[19] Bench experiments might have used computational tools, such as homology modeling or molecular docking software,[20] and computational studies are often externally validated or supplemented by *in vitro* and *in vivo* laboratory experiments.[21] In this review and associated analyses, we focus largely on performance evaluation of computational technologies; however, the metrics discussed herein are
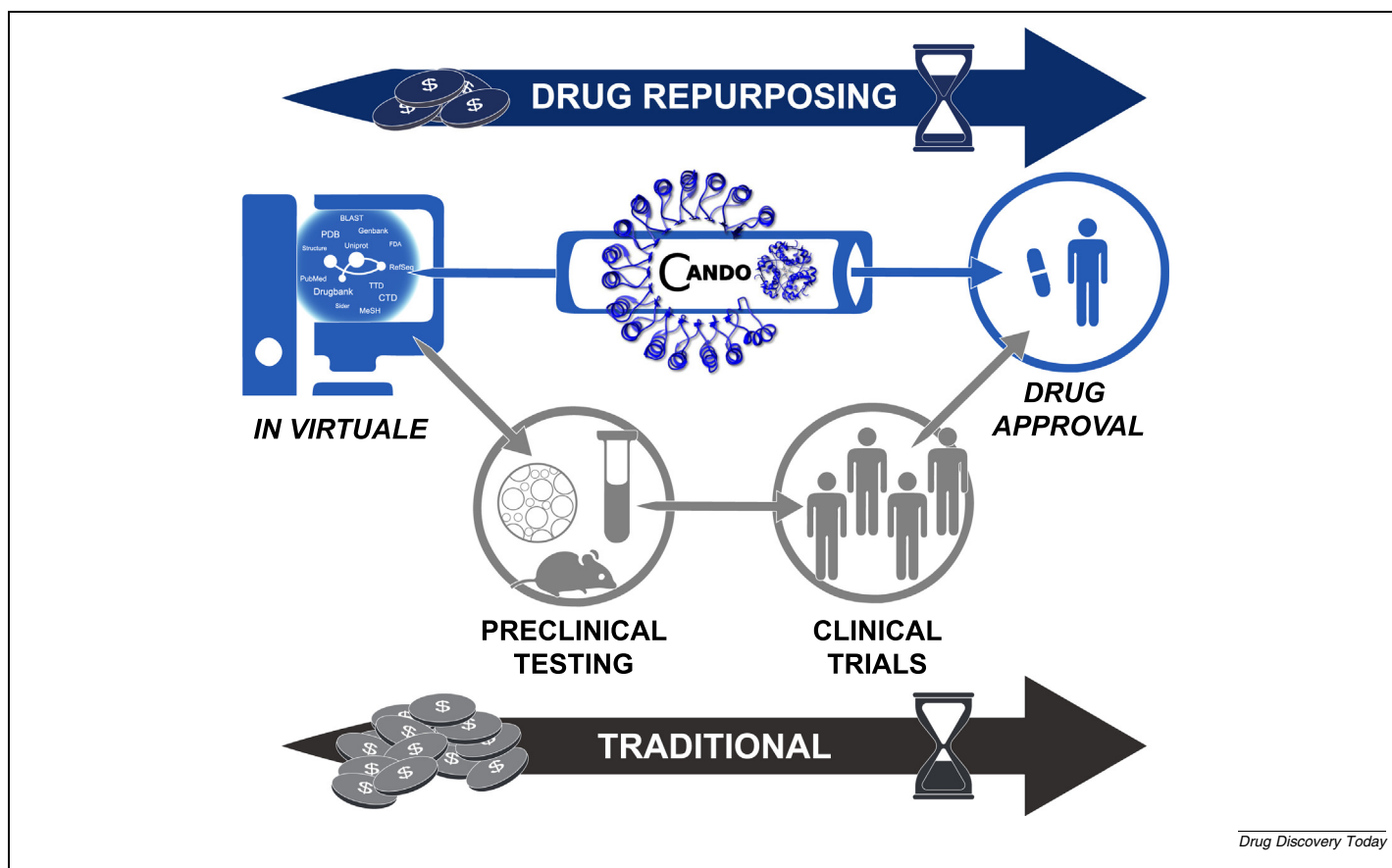

FIGURE 1

The relationship between drug-repurposing technologies and traditional approaches. Traditional drug discovery and development are time consuming and costly, moving from preclinical research (basic computational methods, intensive *in vitro* and/or *in vivo* screening) to testing in clinical trials and eventual drug approval. The ideal scenario is a clinician/physician utilizing results of drug-repurposing technologies [such as Computational Analysis of Novel Drug Opportunities (CANDO)] directly, prescribing medication with high confidence to treat numerous indications, thereby saving time, cost, and improving patient outcomes. In this future guided by using the best evaluation metrics, repurposing technologies have high comparability and fidelity to reality.

applicable broadly to drug repurposing, that is, any technology that generates novel therapeutic repurposing candidates by benchmarking drug–indication association predictions.

Most previous reviews of drug-repurposing technologies have focused on methods development,[4,22–36] with only a few providing cursory analyses of evaluation of those methods.[30,36,37] Brown and Patel previously reported a review of 'validation' strategies for computational drug repurposing,[38] broadly categorizing various evaluation metrics into '(1) validation with a single example or case study of a single disease area, (2) sensitivity-based validation only and (3) both sensitivity- and specificity-based validation'.[38] Shahreza et al., in reviewing network-based approaches to drug repurposing, mention various evaluation criteria used in different studies and provide mathematical aspects of the relationship between some of them.[31]

Here, we augment and enhance their foundational reporting by describing, reviewing, and analyzing metrics for evaluating the performance of drug-repurposing technologies. We highlight uses of metrics borrowed from the realms of virtual screening/target prediction and information retrieval, and report results of their integration into CANDO. Through the use of better evaluation metrics, we aim to make drug-repurposing science more rigorous and comparable. This study will help enable proper evaluation of drug-repurposing technologies, and ultimately guide the field to bring about real changes in the armamentarium of medicine to alleviate disease burden.

CANDO

We developed and deployed the CANDO platform to model the relationships between every disease/indication and every human use drug/compound.[8–18] Built upon the premise of polypharmacology and multitargeting, at the core of CANDO is the ability to infer similarity of compound/drug behavior. Canonically, we use molecular-docking protocols to evaluate the interaction between large libraries of drugs/compounds and protein structures. We then construct a compound–proteome interaction signature to characterize and quantify their behavior. Based on the similarity of these interaction signatures, we rank every drug/compound relative to every other. We hypothesize that drugs/compounds with similar interaction signatures could be repurposed for the same indication(s).

Since the development and application of CANDO version 1,[8–11] we have continued to enhance our platform by analyzing the effect of protein subsets on drug behavior, implementing heterogeneous measures of drug/compound similarity, using multiple molecular-docking software packages to evaluate interactions, and refining nonsimilarity-based approaches for drug repurposing in situations where there is no approved drug for a disease/indication.[12–18,39]

CANDO v2

Version 2 of the CANDO platform (v2) described here, implementing updated drug/compound and protein structure libraries, indication lists, drug–indication mappings, interaction scoring protocols, benchmarking and evaluation metrics, along with data fusion of multiple pipelines mixing and matching between these choices, is used as a template for the rigorous eval-

uation of the performance of drug repurposing technologies. The core tenets remain the same as in CANDO v1; however, the updated data and evaluation metrics enable us to better determine the correctness of those predictions with greater cross-platform comparability, as well as agreement with preclinical and clinical validation experiments.

Curation of drug/compound and protein structure libraries
The v2 compound library contains 2162 US Food and Drug Administration (FDA)-approved drugs extracted from DrugBank 5.0.[40] We have also created a larger library of 8752 compounds from DrugBank containing both approved drugs and experimental/investigational compounds. The updated protein library contains a nonredundant set of 14 606 solved structures compiled from the Protein Data Bank (PDB).[41] Supplementing the approved drugs with compounds that are in the final stages of the drug development process helps to expand the repurposing capabilities of the platform. The updated protein structure library comprises individual chains from numerous proteomes with an equitable distribution of folds and binding sites, allowing for greater coverage because of homology while reducing bias.

Interaction scoring
The default pipeline in CANDO v2 uses an enhanced version of our previous bioinformatic docking protocol for interaction scoring.[9] We altered the previous protocol by using extended connectivity fingerprints from RDKit[42] instead of FP2 fingerprints from OpenBabel[43] as our cheminformatic approach for drug/compound molecular fingerprinting. The other major modification was to use COACH to predict protein structure binding sites.[44] COACH leverages the results from multiple binding-site prediction software suites, including COFACTOR, which we previously used exclusively. The use of the updated fingerprints and COACH results in higher fidelity to reality for the resulting interaction scores from the bioanalytics-based docking (BANDOCK) protocol, as evaluated by comparing results to observed compound–protein interaction binding constants obtained from PDBBind.[13]

Drug/compound characterization, benchmarking, evaluation metrics, and performance
The default pipeline in CANDO v2 generates drug–proteome interaction signatures by calculating scores between all 2162 drugs and all 14 606 proteins using the BANDOCK protocol. Every drug is characterized by a unique vector of interaction scores. We measure the pairwise similarity/distance between the interaction signatures from each of the 2162 drugs to every other. We evaluated a variety of similarity measures between a pair of interaction signatures and found success using the root mean squared deviation (RMSD), which is the default measure, and the cosine distance.[45] We then sort drugs relative to one another according to their similarity (i.e., those with the greatest similarity to each other are ranked highest).

Each drug is associated with one or more approved indications, comprising our standard to which we compare our rankings (also known as a 'gold standard' or 'ground truth'). These associations, or drug–indication mapping, is derived from curating the Comparative Toxicogenomics Database (CTD),[46] which

uses medical subject headings (MeSH) to label indications.[47] This yields 18 709 associations for the 2162 drugs covering a total of 2178 indications. The default benchmarking protocol implements an in-house leave-one-out procedure to accurately identify related compounds approved for the same indication.[9,16] For every indication associated with a drug, we calculate the ranks of other drugs associated with that same indication, and whether any positive hit occurs within certain cutoffs, such as top10, top25, and so on, representing the top 10 and 25 most similar drugs. For each indication, we calculate the percentage of associated drugs that achieve a hit in that cutoff. We next calculate the mean of all per-indication accuracies to give an overall evaluation of the platform, referred to as the average indication accuracy (AIA).[9,16]

For drug repurposing, our small-molecule library is limited to the 2162 approved drugs, but CANDO is capable of analyzing compounds that are not yet approved in a similar fashion. Similar drugs/compounds not associated with the same indication are hypothesized to be novel repurposed therapies to be validated via external preclinical and clinical studies.

Consider our results for the indication melanoma (MeSH ID: D008545), which is associated with a curated list of 58 drugs from the CTD. CANDO predicts 23 of those 58 drugs to have another associated with melanoma within their respective top 10 most-similar proteomic interaction signatures. Therefore, the top10 indication accuracy for melanoma is 23/58 × 100 = 39.6%. We repeat this process for every one of the 2178 indications to generate the AIA.

For CANDO v1, we achieved a top10 AIA of 11.8%, compared with a reported random control of 0.2%.[9] Using improved bio- and cheminformatic tools and armed with a better understanding of random controls, theoretically modeled using a hypergeometric distribution and empirically measured using uniformly random drug–drug similarity data, version 1.5 of CANDO achieved an average indication accuracy of 12.8% at the top 10 cutoff against a random control of 2.2%.[13]

The average indication accuracy is not used by others in the field of drug-repurposing technologies. Thus, although we use it as a metric for internal comparison (i.e., between individual CANDO pipelines and versions), the cross-platform applicability is low. Therefore, we researched other methods of assessing performance, which are now implemented in and applied to CANDO.

Classification, ranking, metrics, and integration into CANDO

Experiments using drug-repurposing technologies return results as a classification or ranking. In classification, compounds are associated with indications in a binary fashion based on some criteria, whereas, in a learning-to-rank experiment, entities are ranked relative to one another in order of some score. Best pairings, designated by a specific rank/classification cutoff/threshold, are reported as putative therapeutics. A ranking result can be thought of as classification by using the cutoff as a threshold for the ranks and declaring items on one side of the threshold as positive samples and those on the other side as negative samples. The inverse of modeling classification as a ranking problem
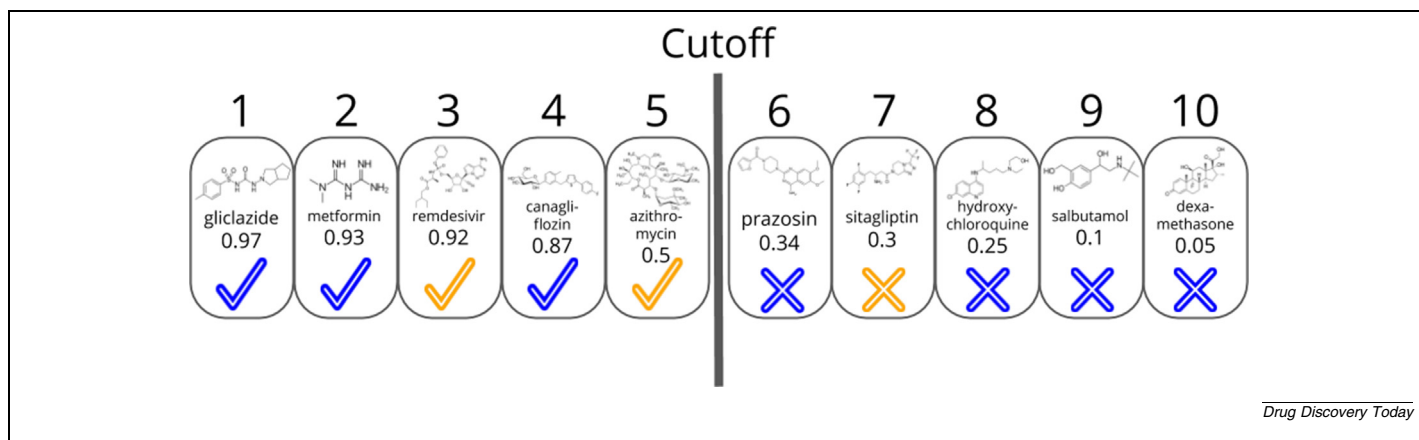

Drug Discovery Today

FIGURE 2

Mock example of novel treatment prediction for type 2 diabetes mellitus (T2DM) by a drug-repurposing technology. We illustrate an arbitrary methodology that ranks ten drugs to treat T2DM, with rank one (number above boxes) being the most likely efficacious and rank ten being the least likely. These ranks are based on mock *scores* given below the name of each drug. Classification of compounds with a score ≥ 0.5 are the same as those ranked in positions one through five (i.e., the results of a classification and ranking schemes are interchangeable). Drugs that have known treatment associations with T2DM are marked with a blue check or cross, and those with unknown associations are marked correspondingly in orange. Those drugs classified/ranked better than the cutoff are the positive results, whereas those worse are the negative results. Given that there are three drugs with a known association to T2DM (blue check) and two with no known association (orange check) ranked better than or equal to rank five, there are three true positives and two false positives. Analogously, there are four true negatives (blue cross) and one false negative (orange cross). If this were not a mock example, the false positive results would be the repurposing candidates for the given indication. We also note that the notion of 'true negative' can be misleading, because most such associations are of undetermined classification, having never been rigorously scientifically studied. This lack of negative data in comparison standards is a limitation in the evaluation of drug-repurposing technologies. This is a mock example of a single ranking among hundreds or thousands in a comprehensive drug repurposing platform. Metrics using a single ranking are averaged over many rankings to produce values that describe the correctness of a drug-repurposing technology.

is also true. If a specific cutoff is used for calculation of performance, it must be reported, and reporting values at all possible ranking and classification cutoffs/thresholds will allow for greater comparability.

An inherent limitation of drug-repurposing technologies and nearly all metrics we use to report on their goodness is the forced dichotomization of results, where those drug–indication associations ranked/classified better than the cutoff/threshold are labeled 'positive' and those worse are 'negative', with no nuance or allowance for real-world considerations, such as first- and second-line therapies. This lowers the fidelity of all drug-repurposing models to reality. Fig. 2 illustrates these issues in a mock example of predicted therapies to treat type 2 diabetes mellitus by a drug-repurposing technology.

Drug-repurposing technologies use a multitude of different metrics to evaluate the performance of the resulting ranking or classification. These are based on delineation of results into true positives (TPs), false negatives (FNs), true negatives (TNs), and false positives (FPs) relative to some standard. Notable commonly used metrics include sensitivity (TP rate and recall), specificity (TN rate), false discovery rate, FP rate precision (positive predictive value), area under the receiver operating characteristic curve (ROC and AUROC), precision, precision-recall curves, and area under precision-recall curves, F1-score, and Matthews correlation coefficient (MCC; see the supplemental information online).

Results of a drug-repurposing experiment comprising a ranking of drug candidates are similar to those results obtained in virtual screening and target prediction experiments, but the standard of comparison is different: known drug characteristics or drug–target associations (binding interactions) as opposed to known drug–indication associations. Drug repurposing as a field is not one-to-one with drug target prediction,[38] but a drug target prediction experiment can be part of a repurposing experiment. Despite this demarcation, metrics used in virtual screening/target prediction highlighting the 'early recognition problem' may be useful to evaluate drug repurposing.[48,49] In both instances, the goal is to prioritize ranking 'active' candidates (known drug–target or drug–indication associations) at the top. Metrics that consider the early recognition problem properly include the enrichment factor (EF; see the supplemental information online), robust initial enhancement (RIE; see the supplemental information online),[50] and the Boltzmann-enhanced discrimination of ROC (BEDROC).[48]

In addition to being similar to virtual screening, the results of drug-repurposing technologies are highly analogous to information retrieval. Information retrieval tools, such as web search engines, can be evaluated in their ability to accurately return a website link that is desired and subsequently visited by the user based on some query string. In a drug-repurposing experiment, this is akin to generating a list of active candidate drugs that might be a treatment for a given indication. Therefore, the goals of information retrieval and drug repurposing are similar, and correspondingly performance metrics that have been explored for information retrieval have value in drug-repurposing evaluation. These include the mean reciprocal rank (MRR), precision-at-K (P@K), average precision (AP) and mean average precision

(MAP; see the supplemental information online), and (normalized) discounted cumulative gain [(N)DCG].[51]

We describe the utility and use of the above-mentioned metrics in several computational drug-repurposing experiments. We report the complete results of using every metric at all cutoffs for both versions 1.5 and v2 of CANDO in the supplemental information online and at our website, and highlight specific results of interest herein. In general, most metrics have use for internal intraplatform comparisons but limited use for external interplatform comparisons. Based on our analyses, we conclude that the metrics with the most utility relative to their cost are BEDROC and NDCG. Additionally, we have integrated many of these measures into the CANDO platform to facilitate internal and external comparability. Evaluation of CANDO using these newly integrated metrics has reaffirmed its utility for drug repurposing, while providing a foundational review of the advantages and limitations of each metric in the context of the libraries and standards used.

Measures of correctness/success
Mean reciprocal rank
Reciprocal rank is the inverse of the position of the first correctly retrieved active in a ranking scheme, or the best-scoring active in a classification. The correctness/success is determined by matching the retrieved active to a known drug–indication association according to some standard. MRR is the average of the inverse rank of each first retrieved active (i.e., the first TP).[52] Although easy to calculate, this metric only uses the ranking of the first retrieved active. According to MRR, a drug-repurposing experiment that ranks a single active correctly out of several performs just as well as another that ranks several correctly. An overall measure of correctness is difficult to discern from the reporting of a single value; however, a possible way to evaluate distributions of performance across several experiments is provided by Eq. (1)[52]:

$$MRR = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{rank_i} \qquad (1)$$

where m is the total number of measurements made and $rank_i$ is the rank of the first active.

MRR is the least similar to the other metrics reviewed. It does have utility ranking putative drug candidates for an indication for which there is a single known association (i.e., there is only one active to compare against for evaluation of correctness/success), such as with some neglected and emerging indications. Our previous metric for evaluating performance in CANDO, AIA, is similar to MRR, in that, for each drug–indication pair, we are primarily concerned with whether there is an active within a certain cutoff.

True positive, true negative, false positive, and false negative
Given a classification or ranking of candidate drugs for repurposing, the positive samples retrieved are those deemed to be associated with the desired indication or correctly ranked within the specified cutoff. Negative samples are those classified as having no association with the desired indication or ranked outside the cutoff. Within these, there are TP, FP, TN, and FN samples. TP samples are correctly classified or ranked associations that

are present in the standard, whereas FPs are incorrectly classified as positive or ranked within the cutoff, with the corresponding association not in the standard. FNs are known associations found in the standard but classified incorrectly or ranked outside the cutoff, whereas TNs are not present in the standard and correctly classified as such or ranked outside the cutoff.

Sensitivity and specificity

Sensitivity is the proportion of TPs that are correctly identified; specificity is the proportion of TNs that are correctly identified (Eq. (2)):

$$sensitivity = recall = true\ positive\ rate\ (TPR) = TP/(TP + FN)$$
$$specificity = selectivity = true\ negative\ rate\ (TNR) = TN/(TN + FP) \tag{2}$$

Lim et al.[53] used sensitivity to report drug–target prediction correctness using their REMAP platform. They did not directly quantify their drug-repurposing predictions, but found corroborating examples of novel treatments in the literature. Donner et al. used recall as a metric for reporting and visualization of ranking perturbagens (chemical substances that change gene expression),[54] and Xuan et al. graphically showed the recall at top cutoffs of rankings of drug–indication associations generated via their methods.[55] Wu and colleagues used sensitivity as one of their metrics of choice to evaluate the ability of their repurposing platform (MD-Miner) to identify active drugs among top-ranked candidates for repurposing.[56] The reporting of sensitivity and
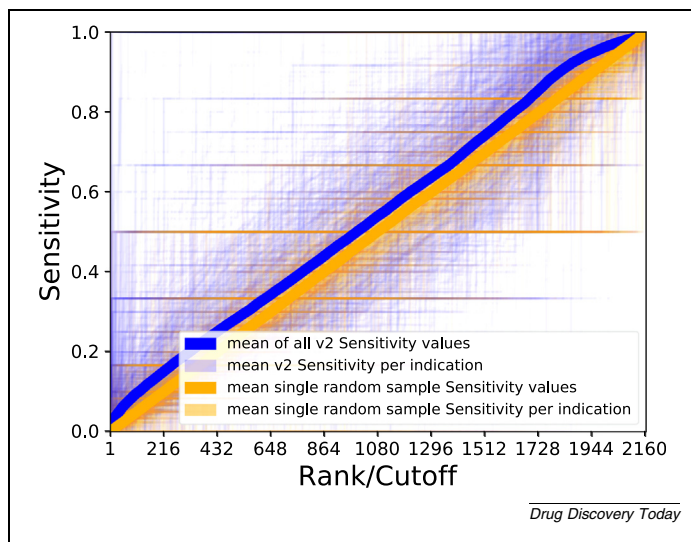

FIGURE 3

Evaluating Computational Analysis of Novel Drug Opportunities (CANDO) performance using sensitivity. The sensitivity values (vertical axis) of all drug–indication associations in CANDO v2 (blue) and a random control (orange) are shown according to the rank (horizontal axis). Broadly, more drug–indication pairs score better at all ranks using the v2 pipeline relative to random, visually observed as more blue in the upper left half and purple in the lower right half above. The darkness of a point is directly proportional to how many lines pass through that point. This is an illustration of results using a single metric for a single pipeline within a single platform compared with the same pipeline with random input data, highlighting the difficulties in visualizing data of this type and size. More metrics, more pipelines, more platforms, and more data (including controls) will greatly increase the complexity of this illustration.

specificity is not usually the focus in these studies. Instead, they are described and displayed as part of another metric, often the receiver operator characteristic curves and area under such curves discussed below. Fig. 3 illustrates the application of this metric to CANDO.

False discovery rate and false positive rate

More TPs are classified and ranked correctly when quantifying and subsequently limiting the number of FPs. In drug repurposing, the false discovery rate (FDR) is typically used as a cutoff for further development of specific results and not as a standalone metric. Through consideration of drugs, inflammatory bowel disease (IBD) genes, and biological pathways, Grenier and Hu generated candidate treatments for IBD, using the FDR as a way to guide classification of putative therapeutics.[57] Sirota et al. used FDR to measure the significance of drug–indication scores compared with random in their experiments comparing gene expression levels as drug signatures.[19] Hingorani et al. claimed that drug discovery projects fail because of their excessive FDR[58]; in addition, they calculated the probability of repurposing success based on several assumptions. A utility of drug-repurposing technologies is to reduce the FDR, but current methods can easily inflate the number of false discoveries by generating excessive numbers of predictions.[59] Lim et al. used a confidence weight to quantify uncertainties in predictions to reduce FPs.[60] In a similar way to the FDR, the false positive rate (FPR) is generally used as an integral part of another metric (Eqs. (3) and (4)).

$$FDR = FP/(FP + TP) \tag{3}$$

$$FPR = FP/(FP + TN) \tag{4}$$

Receiver operating characteristic curve and area under the ROC curve

ROC curves are graphs in which each point is the representation of a binary classification performance measured using the TP and FP rates. The points along an ROC curve are discrete but are often shown as continuous lines that are obtained by varying the cutoff for classification or rank value and calculating the corresponding TPR and FPR. One of the more popular methods for assessing and reporting the performance of drug-repurposing technologies is the area under the ROC curve (AUC or AUROC). A singular value, the AUROC is calculated either using the trapezoid rule or directly using the rank of the actives (drug–indication associations present in the standard) along with the ratio of actives and ratio of inactives (to be determined associations, or not present in the standard) in the entire drug library.[48] We have implemented the second approach in CANDO (Fig. 4). A higher value is taken to be indicative of better performance, with a perfect classification obtaining an AUROC of 1.0, and 0.5 indicating random ranking/classification (Eqs. (5) and (6)):

$$AUROC = \frac{1 - (x)}{R_i} - \frac{R_a}{2R_i} \tag{5}$$

where R_a is the ratio of actives, R_i the ratio of inactives, and

$$x = \frac{1}{nN} \sum_{i=1}^{n} r_i \tag{6}$$
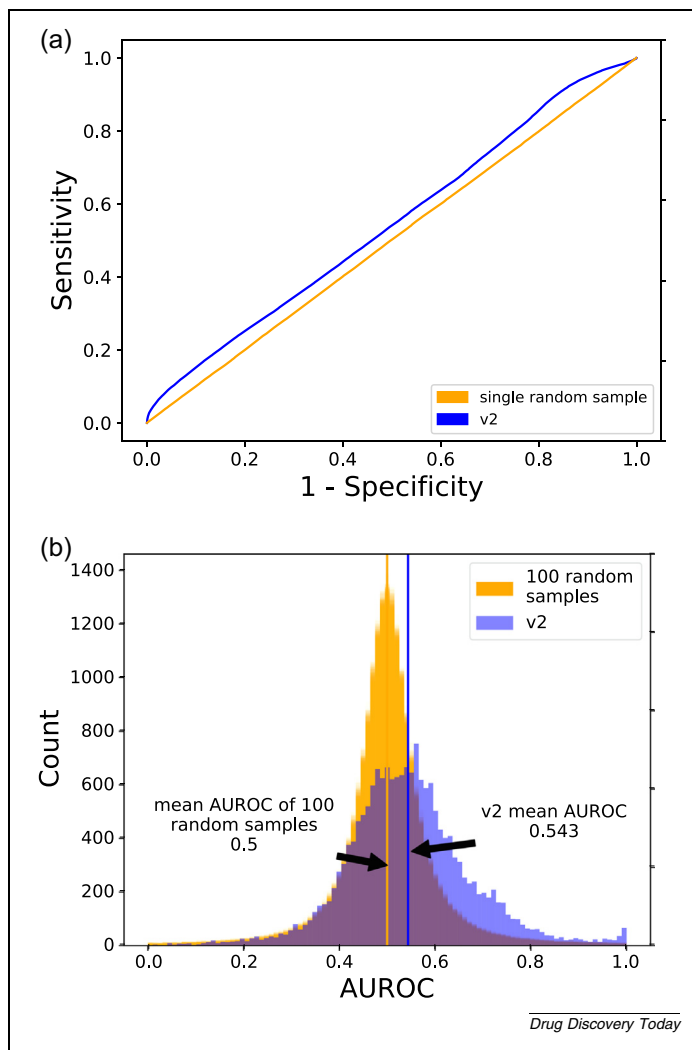
FIGURE 4

Evaluating Computational Analysis of Novel Drug Opportunities (CANDO) performance using receiver operating curves (ROC) and area under ROC (AUROC). **(a)** Mean ROC curve across all drug–indication pairs for CANDO v2 (blue) and a single sample random control (orange). As with all ROC curves, the horizontal axis is 1–specificity and the vertical axis is sensitivity. The empirical random matches what is mathematically expected by random (a straight line along the diagonal) (i.e., both reflecting drug–indication associations obtained by chance based on a uniform distribution). **(b)** Histogram of AUROC scores for all drug–indication pairs predicted by v2 (blue) compared with those from 100 random runs (orange). The mean AUROC of all drug–indication pairs from v2 is 0.543, compared with a empirically derived random mean of 0.5 (again, matching theoretical expectation). The right shift of the v2 AUROC indicates an overall better performance relative to random controls. Both ROC and AUROC are useful for internal validation, but overall have limited utility in assessing different drug-repurposing experiments or technologies, in part because of imbalances in known drug–indication associations and not emphasizing early recognition.

where r_i is the rank of the ith active, and N is the total number of drugs

The drug-repurposing project PREDICT uses AUROC as a main method of reporting goodness.[61] Moridi *et al.* reported their own AUROC compared with that of PREDICT.[62] Nguyen *et al.* created a computational drug-repurposing framework based on control system theory (DeCoST) to make novel treatment predictions

for cancer.[63] Notably, Nguyen *et al.* included negative associations in their studies, such as drugs withdrawn from treatment or terminated clinical trials, which enhanced the fidelity of their computational experiments to reality.[63] Given the different libraries and standards used, drug-repurposing studies are better analyzed with one or more other metrics in addition to the AUROC. As stated previously, the AUROC is dependent on the ratio of actives to inactives in a library; the DeCoST framework overcame this dependency issue in part by creating a new, more balanced, library derived from drugs used by another group.[64] Emre Guney used AUROC as a metric, pointing out how data in a drug-repurposing experiment might bias scientists toward conclusions that are not justified, and how single values of AUROC might not hold up to further cross-validation.[65]

Lee *et al.* compared results of their drug-repurposing experiments across a diverse breadth of indication types using AUROC, showing better performance than those of others for the same indication types.[66] The mean AUROC of CANDO v2, calculated on a per indication basis, which itself is a corresponding average of grouped drug–indication pairs when there is more than one drug, is 0.520, with a median of 0.525 (interquartile range: 0.481–0.561).

The biggest shortcomings of the ROC/AUROC are the lack of early recognition and inability to handle imbalanced data. As illustrated with respect to virtual screening,[48] the metric fails to enable comparison of drug repurposing for ranking actives at the top of an ordered list, which is the desired goal.

Precision and precision-recall, and area under the precision-recall curve

Precision measures the relevance of a set of predictions (Eq. (7)):

$$precision = TP/(TP + FP) \qquad (7)$$

Yu *et al.* use established disease–gene–drug relationships to infer new drug–tissue–specific disease relationships, and reported precision as a standalone metric (as a score relative to the top percent of drug–indication pairs).[67] Precision is often reported alongside recall, and one of the most commonly used metrics used to evaluate drug-repurposing technologies is area under the precision-recall (PR) curve (AUPRC). Saito and Rehmsmeier provide strong evidence for the superiority of precision-recall compared with ROC when evaluating imbalanced data.[68] Imbalanced data are commonplace in drug–indication association standards used by repurposing technologies (i.e., the drugs in the standard are spread divergently across the indications, and vice versa).

PR curves show a distinctive jagged edge pattern or appear finely interpolated, retaining maximum precision up to a particular recall.[51] Xuan *et al.* used both ROC and PR to compare their drug–indication association predictions to those made by others.[55] Peng *et al.* also used ROC and PR curves to demonstrate the internal validation of their network-based inferences about drug Anatomic Therapeutic Classification (ATC) codes.[69] McCusker *et al.* used precision to evaluate their computational drug-repurposing predictions to treat melanoma as 'the percentage of returned candidates that have been validated experimentally or have been in a clinical trial versus all candidates returned'.[70] This metric is indeed precision, albeit based on a different type

KEYNOTE (GREEN)

of standard, because they applied it to evaluate capturing literature instead of known associations.

Iwata *et al.* used AUPRC to assess the internal performance of reconstructing known drug–indication associations made using supervised network inference, and compared their results to those obtained by others.[71] Khalid and Sezerman reported AUPRC along with AUC and mean percentile rank to demonstrate the ability of their platform, which combines protein–protein interaction, pathway, protein-binding site structural and disease similarity data, to capture known drug–indication associations.[72] With respect to cross-platform comparability, the authors applied their algorithm to evaluate performance using gold standard data available online for three other platforms and found that they obtained better AUC values using their methods but with data from other platforms.[72] Similarly, using both AUC and AUPRC, Wang *et al.* used DeepDRK, 'a machine learning framework for deciphering drug response through kernel-based data integration' to predict individualized patient responses to chemotherapy.[73]

Accuracy and F1-score

The term 'accuracy' causes confusion because it is used to refer to performance generally in a colloquial sense and also to a mathematically defined value in the context of binary classification. In a drug-repurposing evaluation context, accuracy is the fraction of TPs and TNs correctly classified (Eq. (8)):

$$accuracy \; = \; (TP + TN)/(TP + FP + TN + FN) \tag{8}$$

Accuracy is influenced by the number of actives and inactives in a set and, therefore, its utility is limited. Accordingly, it is best used with balanced data, which are rare in drug-repurposing technology standards. CANDO v2 obtains an average accuracy over all drug–indication pairs of 0.94. This high value is appealing at first but is useless because it is identical to the mean average accuracy over all pairs for data collected over 100 random sample runs. This is because our standard is greatly skewed in (correctly) representing the known drug–indication associations.

The F1-score (F-score or F-measure), widely used in machine-learning applications, is the harmonic mean of precision and recall (Eq. (9)):

$$F1 - score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$
$$= 2 \times precision \times recall/(precision + recall) \tag{9}$$

By focusing on small-value outliers and mitigating the impact of large ones, the F1-score provides an intuitive measure of correctness when using uneven class sizes, unlike accuracy. Just as precision and recall are calculated at a certain cutoff, the ranks at which the measurement is made, or the score used for classification, should be reported along with the F1-score. Using CANDO v2, we obtain a mean F1-score calculated over all drug–indication pairs at the top 100 cutoff of 0.033, compared with the mean of 100 random samples at the same cutoff of 0.023.

Zhang *et al.* used AUC, precision, recall, and F1-score in evaluating their SLAMS algorithm.[74] Their calculation of recall is not directly related to the other measures,[74] an indication of how different groups measure performance variably. Aliper *et al.* reported results of their deep learning platform for drug repurposing based on transcriptomic data using accuracy and F1-score.[75] Specifically, the authors reported not only averages, but also the performance of their platform across three, five, and 12 specific therapeutic-use categories according to the MeSH classification, and provided putative explanations for differences. McCusker *et al.* also used the F1-score in evaluating performance of their probabilistic knowledge graph platform.[70]

Boltzmann-enhanced discrimination of receiver operating characteristic

The BEDROC metric merges early recognition with the area under the ROC.[48] More formally, the BEDROC metric evaluates the probability that an active ranked by the evaluated method will be found before any other that is derived from a hypothetical exponential probability distribution function with parameter α, where $\alpha R_a \ll 1$ and $\alpha = 0$. In this context, R_a is the ratio of actives in the standard and $1/\alpha$ is 'understood as the fraction of the list where the weight is important'.[48] The authors state that BEDROC should be understood as assessing 'virtual screening usefulness' as opposed to 'improvement over random' (which is what the ROC does). In the context of drug repurposing, this may be interpreted as the probability that a drug predicted to treat an indication is ranked better than a drug that is not (Eq. (10)).

$$BEDROC = \frac{RIE - RIE_{min}}{RIE_{max} - RIE_{min}} \tag{10}$$

RIE is itself another metric known as robust initial enhancement (Eq. (11)). RIE_{min} is the calculated RIE when all actives are at the bottom of a ranked list and RIE_{max} when they are all ranked better than any inactives; x_i is the relative rank of the ith active (i.e., $x_i = {}^{r_i}$), where r_i is the rank of the active, N is the total number of drugs/compounds, n is the number of actives, and ^{1} is 'the fraction of the list where the weight is important

$$RIE = \frac{\frac{1}{n} \sum_{i=1}^{n} e^{-\alpha x_i}}{\frac{1}{N} \left(\frac{1 - e^{-\alpha}}{e^{\frac{\alpha}{N}} - 1} \right)} \tag{11}$$

Several computational studies to repurpose drugs have used BEDROC as a metric, albeit not to evaluate drug-repurposing performance. Specifically, Govindaraj *et al.* used BEDROC to assess the ability of their algorithm to detect protein pockets binding similar ligands.[76] Alberca *et al.* used BEDROC to assess virtual screening of protein–ligand interactions.[77] Jain reported on limitations of BEDROC and other metrics that assess early enrichment from a virtual-screening perspective, stating that they are biased to report elevated values based on the total number of positives and negatives.[78] An example of explicit use of BEDROC in drug repurposing is from Arany *et al.*,[79] who used it (along with AUROC) to evaluate the effectiveness of their methods to produce drug rankings with respect to correct Anatomical Therapeutic Chemical (ATC) codes.[79]

Comparing BEDROC scores across different α values is not advised.[48] For a given drug-repurposing technology, users might seek to predict novel drugs across a multitude of indications with highly variable numbers of associated drugs. If an indication has 200 associated drugs, then α should be low to maintain the $\ll 1$ condition. However, such an α may inappropriately lead to highly variable BEDROC scores for indications with a low number of drugs, because small changes in absolute ranking of these
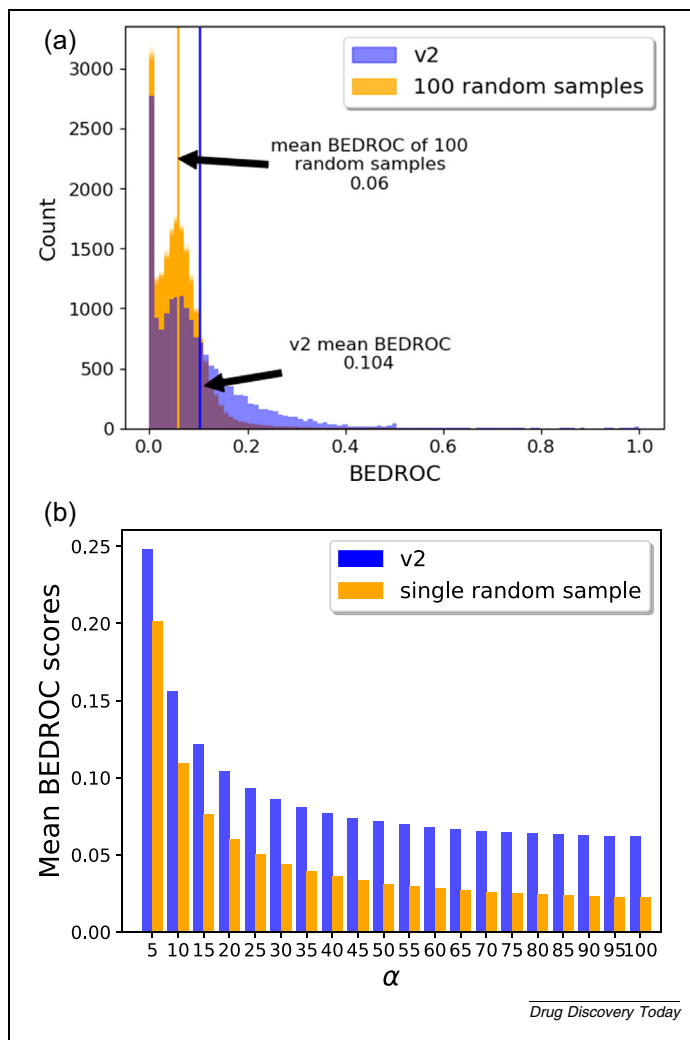
FIGURE 5

Evaluating Computational Analysis of Novel Drug Opportunities (CANDO) performance using Boltzmann-enhanced discrimination receiver operating curve (BEDROC). **(a)** Count of CANDO v2 BEDROC scores using $\alpha = 20$ for all compound–indication pairs in CANDO, compared with 100 samples obtained using shuffled CANDO data (orange). Using $\alpha = 20$, the most commonly used value in the literature, the mean v2 BEDROC score is 0.104, compared with an average of 0.06 over 100 random samples, indicating that v2 outperforms this random control at retrieving known drug–indication associations on average. Generally, BEDROC scores for predicted drug–indication associations from v2 are considerably better than random, indicating their greater real-world utility. **(b)** Mean v2 BEDROC scores (blue) compared with a single random sample (orange) using $\alpha = 5, 10, \ldots, 100$. The random sample was constructed via shuffling of v2 data to obtain drug–drug similarities expected by chance, as usual. At higher values of α, certain drug–indication pairs might violate the conditions necessary for BEDROC to remain useful.

drugs correspond to a large change in the relative ranking. These requirements of α possibly lower cross-platform comparability; regardless, BEDROC is best applied to evaluate repurposing technologies using similar quality and size data (i.e., similar numbers and quality of drugs, indications, and corresponding associations).

Fig. 5 illustrates the results of integrating BEDROC into CANDO. Most uses of BEDROC are based on single α value across

all known drug–indication associations for all calculations to maintain comparability. As platforms become larger and more diverse, finding a balance to handle indications with varying number of associated drugs will be necessary, without being singularly biased to a few well-studied indications. Regardless, BEDROC has several major improvements over AUROC, and the former should be preferred when reporting results of drug-repurposing technologies.

Discounted cumulative gain and normalized discounted cumulative gain

The discounted cumulative gain (DCG; Eq. (12)) is constructed with assumptions that top-ranked results are more likely to be of interest, and that particularly relevant results are more useful.[80] Although data in the form of ranking are readily measured by the DCG, classification schema should be converted to a ranking underpinning the decision boundary to be appropriately measured by DCG. The Ideal DCG (IDCG; Eq. (13)) is the DCG calculated for a ranking in which all known actives are ranked the very best in the prediction list. The Normalized DCG (NDCG; Eq. (14)), with a value between 0 and 1, is obtained by dividing the DCG by the IDCG. The NDCG enables comparison and contrasting of performance evaluation with different numbers of relevant results with meaningful interpretation (i.e., we can use a single value to determine goodness of a drug-indication ranking/classification with greater confidence than most other metrics even when there are different numbers of associations).

$$DCG = \sum_{i=1}^{p} \frac{2^{rel_i - 1}}{log_2(i+1)} \tag{12}$$

$$IDCG = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i - 1}}{log_2(i+1)} \tag{13}$$

$$NDCG = DCG/IDCG \tag{14}$$

where i is the rank of the active in question, up to rank p, and rel_i signifies the relevance of the predicted drug to the indication {0 or 1 in the binary case}, REL_p is the list of associated drugs in the set up to position p, and |REL_p| is the size of that list.

The value of p is a specific position (ranking) at which the NDCG is calculated. Therefore, results are reported as $NDCG_p$. Wang et al. suggest selecting p as a function of the size of the libraries used.[81] The distribution and measures of central tendency of NDCG at a cutoff or multiple cutoffs can be reported (i.e., a NDCG value can be calculated for every possible ranking). One of the most appealing features of DCG is the ability to assign relative importance, captured in the rel_i value. For a certain indication, there might exist more or less effective therapies, which is reflected in the drug–indication association standard. Applied to precision medicine, such a relevance could be determined on a per patient basis. This is a complex variable with a range of possible values but is often used in a binary fashion. The NDCG is the best measure of correctness if a standard has known relative importance assignments.

Ye et al. used NDCG as the metric of choice for analyzing repurposing opportunities based on drug adverse effects.[82] Specifically, the authors reported the top-10 ATC therapeutic categories with $NDCG_5$.[82] Specifically, the authors report the therapeutic categories of drugs (as classified by ATC codes) that
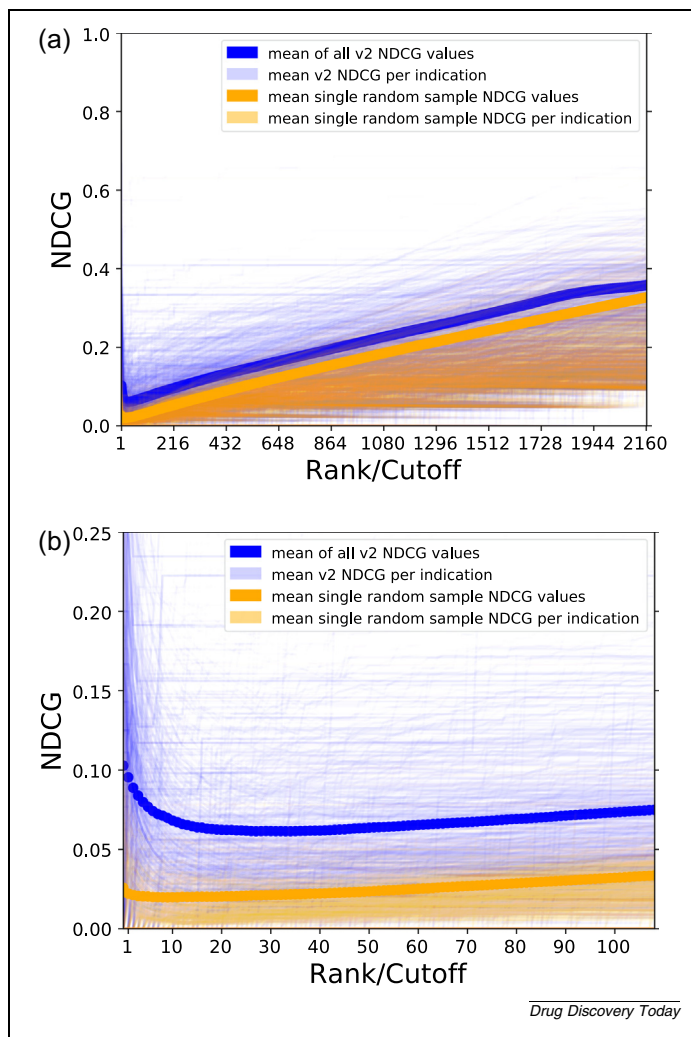
FIGURE 6

Evaluating Computational Analysis of Novel Drug Opportunities (CANDO) performance using normalized discounted cumulative gain (NDCG). The NDCG scores of CANDO v2 compared with those from a single random control at all ranks/cutoffs is shown in **(a)** and at the top 20% of ranks (432) in **(b)**. The overall mean of v2 is the dark-blue line, with per indication means in light blue. The mean of a single random sample is in red, and the per indication means of the same sample is in orange. By chance, some predictions will be worse than random, as is evident whenever an orange line is higher than a light-blue line but, on average, v2 performs better than random at all ranks. These comparisons indicate the utility of CANDO at predicting drug–indication associations using the most rigorous performance evaluation metric considered in this study.

achieved an $NDCG_5$ of more than 0.7 during benchmarking. Although the metric enables comparability, predicting a general categorization of a drug–indication association is easier than a very specific therapeutic prediction. In addition to using NDCG, Ye et al. reported putative therapeutic predictions that overlap with previous similar work.[83] Saberian et al. used NDCG to assess the performance of their drug-repurposing method for three indications [breast cancer, idiopathic pulmonary fibrosis (IPF), and rheumatoid arthritis (RA)], with the score based on the rank of the left-out drug in each round of sampling.[84] Their work included a sample calculation of NDCG in the corresponding

supplementary material, compared with results based on 1000 random rankings of drugs, although the mean NDCG of these rankings was not reported.[84]

Fig. 6 illustrates the mean NDCG at all cutoffs for CANDO v2. We obtained a mean $NDCG_{10}$ of 0.060, compared with an average of 100 random data sets of 0.0197, and a theoretical average of 0.0199. Comparing these values along with the results shown in Fig. 6 indicates that using CANDO to predict drug–indication associations has utility. The elevated cross-platform comparability of NDCG because of the use of logarithmic scoring and normalization makes it among the most-useful metrics reviewed to measure success when used in drug-repurposing technologies (Fig. 7).
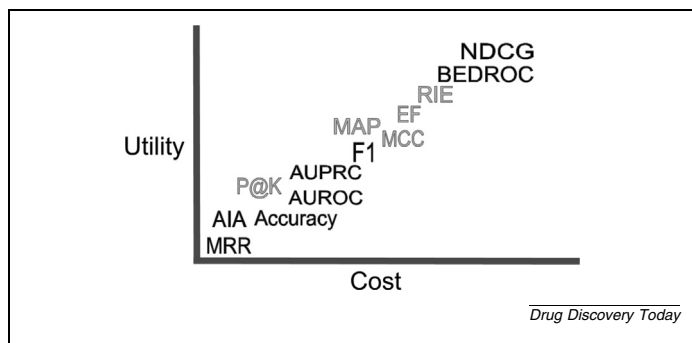
Custom methods of performance evaluation

We have used our AIA metric in CANDO extensively[8–18] (see the section 'Drug/compound characterization, benchmarking, evaluation metrics, and performance'). Similarly, Peyvandipour et al. used a custom evaluation metric in their systematic drug repurposing study.[85] The goal of this review is to more readily compare our results with others, an outcome toward which we are continuously striving, presently by integrating more widely used metrics into our system, and advocating for others to do the same. We initially developed AIA in response to our validation partners seeking a singular successful hit for an indication they were studying; similarly, other researchers might want to use their own evaluation methods for their own reasons. Nonetheless, we recommend researchers also report results using one or more of the metrics described herein.

Evaluating drug repurposing in the context of precision medicine

A specific medication might be more or less efficacious for a particular patient with a specific disease at a given time. Drug-repurposing technologies can be tailored to arbitrary individual contexts and, thus, precision/personalized medicine is a growing area of interest.[33,60,73,86–89] Our group is currently exploring opportunities in the realm of precision cancer therapeutics using both CANDO and our molecular-docking protocol CANDOCK.[17] We have previously published studies to predict HIV drug regimens based on the viral mutations circulating within a patient,[90–92] understanding polymorphisms in the malarial parasite *Plasmodium falciparum*,[93,94] explaining warfarin resistance,[95] among many others. All these studies would have benefited from the application of the evaluation metrics described in this study.

In rare diseases, including rare genetic diseases and rare cancers,[33,96–98] computational drug-repurposing experiments might offer the best chance for discovering efficacious treatments.[4] The field has promising initial results,[86,87,99] but notions of correctness remain limited to mechanistic understanding and preclinical corroboration.[33] The use of particular metrics and quantifiable comparison between experiments is unknown, because it has not been done, but the metrics reviewed herein might have the same utility as when used broadly. Given the low number of individuals with rare diseases, clinical trials are difficult to conduct, and only the most scientifically rigorous preclinical predictions with greatest confidence from drug-

FIGURE 7

A subjective illustration of the relationship between cost and utility of performance evaluation metrics for drug repurposing. Further right on the horizontal cost axis equates to technologies requiring increasing computational power with decreasing intuition. The vertical utility axis is a gauge of enabling cross-platform comparability and, ideally, fidelity to reality through likelihood of success in prospective external validations for multiple drugs per indication. Metrics discussed in the supplemental information online are in gray. We encourage scientists to use metrics such as NDCG and BEDROC, given their high utility and low incremental cost, in addition to avoiding overtraining using a single metric. This study makes the argument for the same general trend to be borne out through prospective clinical validation experiments of drug-repurposing technologies.

repurposing technologies should be considered for further downstream research and use.[96]

Cross-platform comparability and fidelity to reality

Different standards, imbalanced data, and suboptimal design

CANDO was designed to be a shotgun repurposing technology (i.e., to generate putative drug candidates for any/every indication). Lack of consensus with respect to performance evaluation makes it difficult to assess whether this is true of any drug-repurposing technology. A key component of drug-repurposing technologies is the use of some standard drug–indication association library to which results are compared. A 'perfect' evaluation metric cannot correct for data that do not reflect reality, are imbalanced, are over-represented by 'me too' drugs/compounds, or generated as a result of poor cross-validation or overtraining.

There are a limited number of high-quality data sources available for curating drug-repurposing standards, and those that do exist fluctuate significantly. As an example, consider the drug ofloxacin, which is commonly classified as a fluoroquinolone antibiotic. Different standards associate it with between one and 70 indications,[40,46,100–102] which subsequently influences the odds of making a correct prediction by chance. Ultimately, different standards make it difficult to compare technologies and platforms because there are overlaps in semantics in drug–indication associations and language to describe biomedical entities.[103] Moving forward, differences in drug–indication association standards could be overcome through increasing consistency of drug classes across standards,[104] or through integration of ontological understanding into all aspects of drug repurposing,[103] using ontologies specifically designed for this purpose.[105] Knowing the ground truth is a prerequisite for measuring performance,[106] and the use of scientifically rigorous ontologies will ensure robust modeling of reality.

It is natural that there are different numbers of drugs associated with each indication within a particular single standard used in a large platform, such as CANDO, because of biological, economic, and even political reasons. The result is a large discrepancy in the number of drugs approved for, or associated with, a particular indication. For example, in the CANDO v2 drug–indication association standard, there are 218, 216, and 207 drugs associated with pain (MeSH ID: D010146), hypertension (MeSH ID: D006973), and seizures (MeSH ID: D012640), respectively, versus eight for dermatomyositis (MeSH ID: D003882). Although the performance of a drug-repurposing technology platform, such as CANDO, averaged over many indications might be reported as robust, performance on one or a few specific indications is more variable within and between platforms.

Baker *et al.* identified hyperprolific drugs that have been studied in the context of many indications, and indications for which many drugs have been investigated as a treatment.[107] In an attempt to partly overcome variation in results and performance due to chance, Zhang *et al.* eliminated indications with less than ten associated drugs and drugs with less than ten associated indications from their platform.[108] Unfortunately, this action appears to contradict their attempt to meaningfully compare to the PREDICT project using the AUROC, because the distribution of drug–indication associations in PREDICT was enriched in an opposite manner for drugs with less than ten indication associations, and indications with less than ten drug associations.[61]

In describing the evolution of CANDO, we use measures of performance evaluation applied globally, as we have done here. We also apply these measures to specific individual indications, particularly with respect to the prospective validation of the platform or its components.[8,9,11,18,92,109–111] Regarding any classifier technologies, David Hang states, "[A] potential user is not really interested in some 'average performance' over distinct types of data, but really wants to know what will be good for his or her problem, and different people have different problems, with data arising from different domains. A given method may be very poor on most kinds of data, but very good for certain problems".[112] In particular, it is easy to use different input data or comparison standards to obtain numerically better results. Given that a great average performance overall does not guarantee similar performance for specific drugs/indications, and vice versa, drug-repurposing technologies must undergo a thorough vetting across multiple libraries, standards, and experiments (indications to which the technology is applied) to be considered robust.

Inherent limitation of claims and metrics

Metrics for evaluating success of drug repurposing typically rely on the assumption that all associations not part of a standard are negatives. This goes against the entire premise of drug repurposing, which is to expand the list of known drug–indication associations, and any prediction made that is not present in a standard could subsequently be proved correct.

A perfect score for a drug-repurposing experiment based on some evaluation metric does not necessarily mean that perfect

drug-repurposing success was achieved. Numerical evaluation is limited by the choice of cutoff. If the results are an ideal ranking or classification of drugs that are known to be a treatment for an indication against those that are not, then all of the metrics discussed here will yield their best possible result. However, the actionable information is in those drug–indication associations that are truly novel and even unexpected. For example, in a drug-repurposing experiment for breast cancer, the top 10 putative therapeutics reported were known drugs to treat this indication, and the first novel prediction was at rank 11.[113] If any of the metrics described here was used for evaluation with a rank/cutoff of top 10, then this experiment would achieve a perfect score without discovering a novel drug to repurpose.

Purported benefits of high-throughput approaches are the vast size and quick speed, enabling us to explore and make discoveries more quickly than ever before. Thus, specific elements of data/standards used in drug-repurposing experiments and qualitative results (predictions) might sometimes be clinically wrong or nonsensical. This includes incorrect notions of indications[114] and proposition of treatments known to exacerbate disease.[62,115] A few mistakes or inconsistencies in a large drug-repurposing technology do not necessarily invalidate it, but necessitate the need for manual expert inspection and curation.

As artificial intelligence (AI) and machine-learning approaches, including graph-based methods, become more prevalent in drug repurposing,[73,75] the need for rigorous benchmarking has similarly grown. Ascertaining and maintaining the usefulness of these approaches is crucial from both a basic science and clinical perspective. As with other methodologies, biased data as input and overtraining may return results that are falsely interpreted as being significant/high-confidence predictions. Mindset, culture, and willingness to apply computational drug-repurposing models and use their results, considered relevant for the success of AI in drug discovery and development,[116] will partly depend on the confidence in the methods and output, as evidenced by how we evaluate performance. Orthogonal metrics that capture different aspects of the goodness of an experiment could be used in concert to overcome bias and potential voodoo science pitfalls. In the future, prospective blinded assessment of computational drug repurposing, such as those used in protein structure prediction and molecular docking,[117,118] might be another solution to alleviate the problem of bias.

Complex yet quick drug-repurposing technologies can render results beyond the cognitive ability of a person to be familiar with all of its components and the amount of resulting data. The use of rigorous performance evaluation metrics enables a culture in which scientific rigor and correctness are valued more than the novelty in making claims of putative repurposed therapeutics. Even so, it would be prudent for basic science researchers to work with clinicians to ensure that their results make sense to guide correct predictions into clinical use and improve human health.

Validation

In the context of drug-repurposing technologies, 'validation' refers to: internal validation through testing of models on unknown or hidden data; performance as evaluated by the types of metrics discussed here; or to some external independent corroboration. The latter refers to anything from selective reporting of similar results in the literature to results from prospective preclinical and/or clinical studies.

A popular strategy for drug repurposing is to report corroboration of predictions made using computational methods with previously reported independent research in the literature, in a case-based or large-scale analysis.[19,53,119–121] We have used this strategy,[11,12,18] including highlighting literature that contradicts our findings.[15] It is relatively easy with this approach to find examples that support preformed conclusions, and report only those, representing potentially serious instances of confirmation bias. Selective literature corroboration is neither systematic nor hypothesis driven.

Similar to literature analysis is the use of data on clinical trials completed or in progress.[67,74,84,122,123] Through analysis of electronic health records, preventative associations between drugs and indications (i.e., a form of drug repurposing) have been discovered in an ad hoc manner.[124,125] Cheng et al. reported causal increased or decreased risk of coronary artery disease of several drugs based on a networked-based approach to drug repurposing, followed by testing of their predictions using 'large healthcare databases with over 220 million patients' and 'pharmacoepidemiologic analyses'.[126]

There are several examples of preclinical (in vitro and in vivo) validation done following a computational drug-repurposing experiment.[19,21,60,126] In the future, some of these technologies might approach or even rival the current best method for elucidating the usefulness of a drug, which for now remains double-blinded, placebo-controlled, randomized trials with clinically relevant primary endpoints (prolonged life or improved quality of life), and representative samples of subjects, to evaluate both efficacy and safety.[127]

The goal of achieving successful drug repurposing, from technology to clinic (Fig. 1), is still mostly aspirational at this stage. However, progress is being made. The most successful discoveries made using drug-repurposing technologies are ad hoc singular events, which current metrics are not well suited to evaluate.[128,129] Although emotionally unsatisfying, we can search for and use metrics that enable us to compare our technologies directly to each other, for the sake of rigorous science, intellectual merit, and broader impact.

Response to pandemics and novel disease

The potential for drug-repurposing technologies to help respond to epidemics and pandemics rapidly, side stepping lengthy, costly preclinical and clinical studies, is enormous. Recent examples include the Ebola virus disease West African outbreak of 2014, the emergence of the Zika virus, and the COVID-19 global pandemic caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In all instances, there were no drugs approved to treat these indications, but drug-repurposing technologies were used to generate putative therapeutics quickly.[11,18,130–135] It is challenging to use the metrics we have described to evaluate the predictions because there are no previously approved treatments. However, if a platform or methodology has reported measurements of success, especially in related

indications to prevent, treat, or cure infectious diseases, then relying on those performance values as a quantified surrogate could have utility.

To briefly expand on our own work in this area, we applied the CANDO platform at the onset of the COVID-19 pandemic to identify potential small-molecule treatments that inhibit SARS-CoV-2.[18] We first optimized our interaction scoring protocol (BANDOCK) for all compounds against the SARS-CoV proteome via the NDCG metric based on their ability to highly rank a subset of compounds known to be active against various coronaviruses previously identified in two high-throughput screens by Dyall et al.[136] and Shen et al.[137] We then applied the optimized parameter set to the SARS-CoV-2 proteome. We published the top-scoring candidates alongside two other prediction lists generated using our proteomic interaction similarity approach with the v2 human proteome. Currently, 53 out of the 276 approved drug candidates in the three lists have been validated in vitro, in vivo, or in a clinical setting in the literature, including dozens of untested compounds with unknown activity, providing strong evidence that a multitarget approach to discovering repurposed therapeutics is efficient and invaluable, especially when urgent challenges, such as the COVID-19 pandemic, emerge. Continuously updated results are available at http://protinfo.org/cando/results/covid19/ and a full analysis is forthcoming.

The COVID-19 pandemic also illustrates a potential downside of quickly available drug-repurposing predictions. Drugs that have been predicted to be efficacious in treating a known indication might have serious and/or unknown adverse effect profiles when used to treat novel indications. This could be addressed via drug-repurposing technologies by using adverse drug reactions in lieu of indications and similarly benchmarking performance as described here. Studies of rigorous evaluation of drug-repurposing platforms expressed in clear and precise language will help scientists, healthcare workers, institutional and government officials, and the public make informed judgements with respect to future steps on how to use the generated drug candidates for a given indication.

Concluding remarks

Drug repurposing will help advance and evolve therapeutic discovery in the 21st century, bringing new medicines to patients in need. Advancing the field depends on whether we are able to rigorously evaluate the validity and meaning of our computational repurposing experiments with confidence, a crucial component of platform development. We have shown how integration of disparate metrics into the CANDO platform supports this claim. The metrics currently used for gauging correctness of drug-repurposing technologies vary in terms of enabling cross-platform comparability, as well as eventual clinical use of predicted therapeutics. The development and use of improved evaluation metrics will enhance cross-technology comparability and enable more accurate modeling of reality to deliver on the potential of drug repurposing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

References

[1] J. Langedijk, A.K. Mantel-Teeuwisse, D.S. Slijkerman, M.H.D. Schutjens, Drug repositioning and repurposing: terminology and definitions in literature, Drug Discov Today 20 (8) (2015) 1027–1034.

[2] T.T. Ashburn, K.B. Thor, Drug repositioning: identifying and developing new uses for existing drugs, Nat Rev Drug Discov 3 (8) (2004) 673–683.

[3] R. Würth, S. Thellung, A. Bajetto, M. Mazzanti, T. Florio, F. Barbieri, Drug-repositioning opportunities for cancer therapy: novel molecular targets for known compounds, Drug Discov Today 21 (1) (2016) 190–199.

[4] S. Pushpakom, F. Iorio, P.A. Eyers, K.J. Escott, S. Hopper, A. Wells, et al., Drug repurposing: progress, challenges and recommendations, Nat Rev Drug Discov 18 (1) (2019) 41–58.

[5] A.G. Reaume, Drug repurposing through nonhypothesis driven phenotypic screening, Drug Discov Today Therapeutic Strategies 8 (3–4) (2011) 85–88.

[6] Y. Cha, T. Erez, I.J. Reynolds, D. Kumar, J. Ross, G. Koytiger, et al., Drug repurposing from the perspective of pharmaceutical companies, Br J Pharmacol 175 (2) (2018) 168–180.

[7] T.I. Oprea, J.E. Bauman, C.G. Bologa, T. Buranda, A. Chigaev, B.S. Edwards, et al., Drug repurposing from an academic perspective, Drug Discov Today Therapeutic Strategies 8 (3–4) (2011) 61–69.

[8] M. Minie, G. Chopra, G. Sethi, J. Horst, G. White, A. Roy, et al., CANDO and the infinite drug discovery frontier, Drug Discov Today 19 (9) (2014) 1353–1363.

[9] G. Sethi, G. Chopra, R. Samudrala, Multiscale modelling of relationships between protein classes and drug behavior across all diseases using the CANDO platform, Mini Rev Med Chem 15 (8) (2015) 705–717.

[10] G. Chopra, R. Samudrala, Exploring polypharmacology in drug discovery and repurposing using the CANDO platform, Curr Pharm Des 22 (21) (2016) 3109–3123.

[11] G. Chopra, S. Kaushik, P.L. Elkin, R. Samudrala, Combating ebola with repurposed therapeutics using the CANDO platform, Molecules 21 (12) (2016) 1537.

[12] W. Mangione, R. Samudrala, Identifying protein features responsible for improved drug repurposing accuracies using the CANDO platform: Implications for drug design, Molecules 24 (1) (2019) 167.

[13] Z. Falls, W. Mangione, J. Schuler, R. Samudrala, Exploration of interaction scoring criteria in the CANDO platform, BMC Res Notes 12 (1) (2019) 318.

[14] J. Fine, R. Lackner, R. Samudrala, G. Chopra, Computational chemoproteomics to understand the role of selected psychoactives in treating mental health indications, Sci Rep 9 (1) (2019) 1–15.

[15] J. Schuler, R. Samudrala, Fingerprinting CANDO: increased accuracy with structure-and ligand-based shotgun drug repurposing, ACS Omega 4 (17) (2019) 17393–17403.

[16] W. Mangione, Z. Falls, G. Chopra, R. Samudrala, CANDO.PY: Open source software for analysing large scale drug-protein-disease data, bioRxiv 2019 (2019) 845545.

[17] J. Fine, J. Konc, R. Samudrala, G. Chopra, CANDOCK: chemical atomic network-based hierarchical flexible docking algorithm using generalized statistical potentials, J Chem Inf Model 60 (3) (2020) 1509–1527.

[18] W. Mangione, Z. Falls, T. Melendy, G. Chopra, R. Samudrala, Shotgun drug repurposing biotechnology to tackle epidemics and pandemics, Drug Discov Today 25 (7) (2020) 1126–1128.

[19] M. Sirota, J.T. Dudley, J. Kim, A.P. Chiang, A.A. Morgan, A. Sweet-Cordero, et al., Discovery and preclinical validation of drug indications using compendia of public gene expression data, Sci Transl Med 3 (96) (2011) 96ra77.

[20] L. Pinzi, G. Rastelli, Molecular docking: shifting paradigms in drug discovery, Int J Mol Sci 20 (18) (2019) 4331.

[21] B. Chen, L. Ma, H. Paik, M. Sirota, W. Wei, M.S. Chua, et al., Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets, Nat Commun 8 (1) (2017) 1–12.

[22] M.R. Hurle, L. Yang, Q. Xie, D.K. Rajpal, P. Sanseau, P. Agarwal, Computational drug repositioning: from data to therapeutics, Clin Pharmacol Ther 93 (4) (2013) 335–341.

[23] Z. Liu, H. Fang, K. Reagan, X. Xu, D.L. Mendrick, W. Slikker Jr, et al., In silico drug repositioning–what we need to know, Drug Discov Today 18 (3–4) (2013) 110–115.

[24] G. Jin, S.T.C. Wong, Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines, Drug Discov Today 19 (5) (2014) 637–644.

[25] M. Jiao, G. Liu, Y. Xue, C. Ding, Computational drug repositioning for cancer therapeutics, Curr Top Med Chem 15 (8) (2015) 767–775.

[26] J. Li, S.i. Zheng, B. Chen, A.J. Butte, S.J. Swamidass, Z. Lu, A survey of current trends in computational drug repositioning, Brief Bioinform 17 (1) (2016) 2–12.

[27] W. Sun, P.E. Sanderson, W. Zheng, Drug combination therapy increases successful drug repositioning, Drug Discov Today 21 (7) (2016) 1189–1195.

[28] P. Sun, J. Guo, R. Winnenburg, J. Baumbach, Drug repurposing by integrated literature mining and drug– gene–disease triangulation, Drug Discov Today 22 (4) (2017) 615–619.

[29] E. March-Vila, L. Pinzi, N. Sturm, A. Tinivella, O. Engkvist, H. Chen, et al., On the integration of in silico drug design methods for drug repurposing, Front Pharmacol 8 (2017) 298.

[30] Q. Vanhaelen, P. Mamoshina, A.M. Aliper, A. Artemov, K. Lezhnina, I. Ozerov, et al., Design of efficient computational workflows for in silico drug repurposing, Drug Discov Today 22 (2) (2017) 210–222.

[31] M. Lotfi Shahreza, N. Ghadiri, S.R. Mousavi, J. Varshosaz, J.R. Green, A review of network-based approaches to drug repositioning, Brief Bioinform 19 (5) (2017 02;) 878–892.

[32] J.K. Yella, S. Yaddanapudi, Y. Wang, A.G. Jegga, Changing trends in computational drug repositioning, Pharmaceuticals 11 (2) (2018) 57.

[33] B. Delavan, R. Roberts, R. Huang, W. Bao, W. Tong, Z. Liu, Computational drug repositioning for rare diseases in the era of precision medicine, Drug Discov Today 23 (2) (2018) 382–394.

[34] J.M. Pulley, J.P. Rhoads, R.N. Jerome, A.P. Challa, K.B. Erreger, M.M. Joly, et al., Using what we already have: uncovering new drug repurposing strategies in existing omics data, Annu Rev Pharmacol Toxicol 60 (1) (2020) 333–352.

[35] K. Savva, M. Zachariou, A. Oulas, G. Minadakis, K. Sokratous, N. Dietis, et al., Computational drug repurposing for neurodegenerative diseases, in: K. Roy (Ed.), In silico drug design, Elsevier, Amsterdam, 2019, pp. 85–118.

[36] S.S. Sadeghi, M.R. Keyvanpour, An analytical review of computational drug repurposing, IEEE/ACM Trans Comput Biol Bioinform 18 (2) (2021) 472–488.

[37] H. Luo, M. Li, M. Yang, F.X. Wu, Y. Li, J. Wang, Biomedical data and computational models for drug repositioning: a comprehensive review, Brief Bioinform 22 (2021) 1604–1619.

[38] A.S. Brown, C.J. Patel, A review of validation strategies for computational drug repositioning, Brief Bioinform 19 (1) (2018) 174–177.

[39] M.L. Hudson, R. Samudrala, Multiscale virtual screening optimization for shotgun drug repurposing using the CANDO platform, Molecules 26 (2021) 2581.

[40] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, et al., DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res 46 (D1) (2018) D1074–D1082.

[41] H. Berman, K. Henrick, H. Nakamura, Announcing the worldwide protein data bank, Nat Struct Mol Biol 10 (12) (2003) 980.

[42] Landrum G. RDKit: Open-source cheminformatics. https://www.rdkit.org/ [accessed August 6, 2021].

[43] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: an open chemical toolbox, J Cheminform 3 (1) (2011) 33.

[44] Q. Wu, Z. Peng, Y. Zhang, J. Yang, COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking, Nucleic Acids Res 46 (W1) (2018) W438–W442.

[45] A. Singhal, Modern information retrieval: a brief overview, IEEE Data Eng Bull 24 (4) (2001) 35–43.

[46] A.P. Davis, C.J. Grondin, R.J. Johnson, D. Sciaky, R. McMorran, J. Wiegers, et al., The comparative toxicogenomics database: update 2019, Nucleic Acids Res 47 (D1) (2019) D948–D954.

[47] C.E. Lipscomb, Medical subject headings (MeSH), Bull Med Library Assoc 88 (3) (2000) 265.

[48] J.F. Truchon, C.I. Bayly, Evaluating virtual screening methods: good and bad metrics for the 'early recognition' problem, J Chem Inf Model 47 (2) (2007) 488–508.

[49] J. Kirchmair, P. Markt, S. Distinto, G. Wolber, T. Langer, Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes?, J Comput Aided Mol Des 22 (3–4) (2008) 213–228

[50] R.P. Sheridan, S.B. Singh, E.M. Fluder, S.K. Kearsley, Protocols for bridging the peptide to nonpeptide gap in topological similarity searches, J Chem Inf Comp Sci 41 (5) (2001) 1395–1406.

[51] C.D. Manning, P. Raghavan, H. Schutze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, 2008.

[52] E.M. Voorhees, The TREC question answering track, Nat Lang Eng 7 (4) (2001) 361–378.

[53] H. Lim, A. Poleksic, Y. Yao, H. Tong, D. He, L. Zhuang, et al., Large-scale off-target identification using fast and accurate dual regularized one-class collaborative filtering and its application to drug repurposing, PLoS Comput Biol 12 (10) (2016) e1005135.

[54] Y. Donner, S. Kazmierczak, K. Fortney, Drug repurposing using deep embeddings of gene expression profiles, Mol Pharm 15 (10) (2018) 4314–4325.

[55] P. Xuan, Y. Cao, T. Zhang, X. Wang, S. Pan, T. Shen, Drug repositioning through integration of prior knowledge and projections of drugs and diseases, Bioinformatics 35 (20) (2019) 4108–4119.

[56] H. Wu, E. Miller, D. Wijegunawardana, K. Regan, P.R. Payne, F. Li, MD-Miner: a network-based approach for personalized drug repositioning, BMC Syst Biol 11 (5) (2017) 86.

[57] L. Grenier, P. Hu, Computational drug repurposing for inflammatory bowel disease using genetic information, Comp Struct Biotech J 17 (2019) 127–135.

[58] A.D. Hingorani, V. Kuan, C. Finan, F.A. Kruger, A. Gaulton, S. Chopade, et al., Improving the odds of drug development success through human genomics: modelling study, Sci Rep 9 (1) (2019) 1–25.

[59] D. Cavalla, Predictive methods in drug repurposing: gold mine or just a bigger haystack?, Drug Discov Today 18 (11–12) (2013) 523–532

[60] H. Lim, Y.Q. Di He, P. Krawczuk, X. Sun, L. Xie, Rational discovery of dual-indication multi-target PDE/Kinase inhibitor for precision anti-cancer therapy using structural systems pharmacology, PLoS Comput Biol 15 (6) (2019).

[61] A. Gottlieb, G.Y. Stein, E. Ruppin, R. Sharan, PREDICT: a method for inferring novel drug indications with application to personalized medicine, Mol Syst Biol 7 (1) (2011) 496.

[62] M. Moridi, M. Ghadirinia, A. Sharifi-Zarchi, F. Zare-Mirakabad, The assessment of efficient representation of drug features using deep learning for drug repositioning, BMC Bioinform 20 (1) (2019) 577.

[63] T.M. Nguyen, S.A. Muhammad, S. Ibrahim, L. Ma, J. Guo, B. Bai, et al., DeCoST: a new approach in drug repurposing from control system theory, Front Pharmacol 9 (2018) 583.

[64] A. Subramanian, R. Narayan, S.M. Corsello, D.D. Peck, T.E. Natoli, X. Lu, et al., A next generation connectivity map: L1000 platform and the first 1,000,000 profiles, Cell 171 (6) (2017) 1437–1452.e17.

[65] E. Guney, Reproducible drug repurposing: when similarity does not suffice, Pac Symp Biocomput 22 (2017) 132–143.

[66] T. Lee, Y. Yoon, Drug repositioning using drug-disease vectors based on an integrated network, BMC Bioinform 19 (1) (2018) 446.

[67] L. Yu, J. Zhao, L. Gao, Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome, Artif Intell Med 77 (2017) 53–63.

[68] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLoS ONE 10 (2015) e0118432.

[69] Y. Peng, M. Wang, Y. Xu, Z. Wu, J. Wang, C. Zhang, et al., Drug repositioning by prediction of drug's anatomical therapeutic chemical code via network-based inference approaches, Brief Bioinform 22 (2021) 2058–2072.

[70] J.P. McCusker, M. Dumontier, R. Yan, S. He, J.S. Dordick, D.L. McGuinness, Finding melanoma drugs through a probabilistic knowledge graph, PeerJ Comput Sci 3 (2017) e106.

[71] H. Iwata, R. Sawada, S. Mizutani, Y. Yamanishi, Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data, J Chem Inf Model 55 (2) (2015) 446–459.

[72] Z. Khalid, O.U. Sezerman, Computational drug repurposing to predict approved and novel drug-disease associations, J Mol Graph Model 85 (2018) 91–96.

[73] Y. Wang, Y. Yang, S. Chen, J. Wang, DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration, Brief Bioinform (2021), https://doi.org/10.1093/bib/bbab048. Published online April 5, 2021.

[74] P. Zhang, P. Agarwal, Z. Obradovic, Computational drug repositioning by ranking and integrating multiple data sources, in: Joint European conference on machine learning and knowledge discovery in databases, Springer, New York, 2013, pp. 579–594.

[75] A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina, A. Zhavoronkov, Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data, Mol Pharm 13 (7) (2016) 2524–2530.

[76] R.G. Govindaraj, M. Naderi, M. Singha, J. Lemoine, M. Brylinski, Large-scale computational drug repositioning to find treatments for rare diseases, NPJ Syst Biol Appl 4 (1) (2018) 1–10.

[77] Alberca LN, Chugunransky SR, Alv'arez CL, Talevi A, Salas-Sarduy E. In silico guided drug repurposing: discovery of new competitive and noncompetitive inhibitors of falcipain-2. Front Chem. 2019;7:534.

[78] A.N. Jain, Bias, reporting, and sharing: computational evaluations of docking methods, J Comput Aided Mol Des 22 (3–4) (2008) 201–212.

[79] Arany A, Bolg'ar B, Balogh B, Antal P, M'atyus P. Multi-aspect candidates for repositioning: data fusion methods using heterogeneous information sources. Curr Med Chem 2013;20(1):95–107.

[80] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans Inf Syst (TOIS) 20 (4) (2002) 422–446.

[81] Y. Wang, L. Wang, Y. Li, D. He, T.Y. Liu, A theoretical analysis of NDCG type ranking measures, Proc Machine Learning Res 30 (2013) 25–54.

[82] H. Ye, Q. Liu, J. Wei, Construction of drug network based on side effects and its application for drug repositioning, PLoS ONE 9 (2) (2014) e87864.

[83] L. Yang, P. Agarwal, Systematic drug repositioning based on clinical side-effects, PLoS ONE 6 (12) (2011) e28025.

[84] N. Saberian, A. Peyvandipour, S. Donato, S. Ansari, S. Draghici, A new computational drug repurposing method using established disease–drug pair knowledge, Bioinformatics 35 (19) (2019) 3672–3678.

[85] A. Peyvandipour, N. Saberian, A. Shafi, M. Donato, S. Draghici, A novel computational approach for drug repurposing using systems biology, Bioinformatics 34 (16) (2018) 2817–2825.

[86] A. Talevi, Drug repositioning: current approaches and their implications in the precision medicine era, Expert Rev Precision Med Drug Develop 3 (1) (2018) 49–61.

[87] F. Vitali, J. Berghout, J. Fan, J. Li, Q. Li, H. Li, et al., Precision drug repurposing via convergent eQTL-based molecules and pathway targeting independent disease-associated polymorphisms, Pac Symp Biocomput 24 (2019) 308–319.

[88] J.C. Denny, S.L. Van Driest, W.-Q. Wei, D.M. Roden, The influence of big (clinical) data and genomics on precision medicine and drug development, Clin Pharmacol Ther 103 (3) (2018) 409–418.

[89] F. Cheng, W. Lu, C. Liu, J. Fang, Y. Hou, D.E. Handy, et al., A genome-wide positioning systems network algorithm for in silico drug repurposing, Nat Commun 10 (1) (2019) 1–14.

[90] E. Jenwitheesuk, R. Samudrala, Identifying inhibitors of the SARS coronavirus proteinase, Bioorg Med Chem Lett 13 (22) (2003) 3989–3992.

[91] E. Jenwitheesuk, K. Wang, J.E. Mittler, R. Samudrala, Improved accuracy of HIV-1 genotypic susceptibility interpretation using a consensus approach, AIDS 18 (13) (2004) 1858–1859.

[92] E. Jenwitheesuk, K. Wang, J.E. Mittler, R. Samudrala, PIRSpred: a web server for reliable HIV-1 protein- inhibitor resistance/susceptibility prediction, Trends Microbiol 13 (4) (2005) 150–151.

[93] D.-G. Howell, R. Samudrala, J.D. Smith, Disguising itself—insights into *Plasmodium falciparum* binding and immune evasion from the DBL crystal structure, Mol Biochem Parasitol 148 (1) (2006) 1–9.

[94] J. Bockhorst, F. Lu, J.H. Janes, J. Keebler, B. Gamain, P. Awadalla, et al., Structural polymorphism and diversifying selection on the pregnancy malaria vaccine candidate VAR2CSA, Mol Biochem Parasitol 155 (2) (2007) 103–112.

[95] P. Lertkiatmongkol, A. Assawamakin, G. White, G. Chopra, P. Rongnoparut, R. Samudrala, et al., Distal effect of amino acid substitutions in CYP2C9 polymorphic variants causes differences in interatomic interactions against (S)-warfarin, PLoS One 8 (9) (2013) e74053.

[96] P. Pantziarka, L. Meheus, Omics-driven drug repurposing as a source of innovative therapies in rare cancers, Expert Opinion Orphan Drugs 6 (9) (2018) 513–517.

[97] Dhara R. Computational Drug Repurposing for Breast Cancer Subtypes. Univ. of Windsor Masters Thesis; 2019.

[98] S. Ekins, A.J. Williams, M.D. Krasowski, J.S. Freundlich, In silico repositioning of approved drugs for rare and neglected diseases, Drug Discov Today 16 (7–8) (2011) 298–310.

[99] Y.Y. Li, S.J. Jones, Drug repositioning for personalized medicine, Genome Med 4 (3) (2012) 27.

[100] Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z. Wang, et al., Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics, Nucleic Acids Res 48 (D1) (2020 Jan) D1031–D1041.

[101] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, et al., PubChem 2019 update: improved access to chemical data, Nucleic Acids Res 47 (D1) (2019) D1102–D1109.

[102] A.S. Brown, C.J. Patel, A standard database for drug repositioning, Sci Data 4 (1) (2017) 1–7.

[103] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nat Biotechnol 25 (11) (2007) 1251–1255.

[104] R. Winnenburg, O. Bodenreider, A framework for assessing the consistency of drug classes across sources, J Biomed Semantics 5 (1) (2014) 30.

[105] J. Schuler, W. Mangione, R. Samudrala, W. Ceusters, Foundations for a realism-based drug repurposing ontology, in: *Proceedings of the 10th ICBO. City; Publisher*, 2019, p. 7.

[106] C.E. Metz, Basic principles of ROC analysis, Semin Nucl Med 8 (4) (1978) 283–298.

[107] N.C. Baker, S. Ekins, A.J. Williams, A. Tropsha, A bibliometric review of drug repurposing, Drug Discov Today 23 (3) (2018) 661–672.

[108] W. Zhang, X. Yue, W. Lin, W. Wu, R. Liu, F. Huang, et al., Predicting drug-disease associations by using similarity constrained matrix factorization, BMC Bioinform 19 (1) (2018) 1–12.

[109] E. Jenwitheesuk, R. Samudrala, Identification of potential multitarget antimalarial drugs, JAMA 294 (12) (2005) 1487–1491.

[110] E. Jenwitheesuk, J. Horst, K. Rivas, W. Vanvoorhis, R. Samudrala, Novel paradigms for drug discovery: computational multitarget screening, Trends Pharmacol Sci 29 (2) (2008) 62–71.

[111] J.M. Costin, E. Jenwitheesuk, S.M. Lok, E. Hunsperger, K.A. Conrads, K.A. Fontaine, et al., Structural optimization and de novo design of dengue virus entry inhibitory peptides, PLoS Negl Trop Dis 4 (6) (2010) e721.

[112] D.J. Hand, Classifier technology and the illusion of progress, Stat Sci 1–14 (2006).

[113] E. Jadamba, M. Shin, A systematic framework for drug repositioning from integrated omics and drug phenotype profiles using pathway-drug network, BioMed Res Int 2016 (2016).

[114] M. Rastegar-Mojarad, Z. Ye, J.M. Kolesar, S.J. Hebbring, S.M. Lin, Opportunities for drug repositioning from phenome-wide association studies, Nat Biotechnol 33 (4) (2015) 342–345.

[115] L. Cheng, J. Li, P. Ju, J. Peng, Y. Wang, SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association, PLoS ONE 9 (6) (2014) e99415.

[116] P. Schneider, W.P. Walters, A.T. Plowright, N. Sieroka, J. Listgarten, R.A. Goodnow, et al., Rethinking drug design in the artificial intelligence era, Nat Rev Drug Discov 1–12 (2019).

[117] J. Janin, K. Henrick, J. Moult, L.T. Eyck, M.J. Sternberg, S. Vajda, et al., CAPRI: a critical assessment of predicted interactions, Proteins: Struct, Funct, Bioinform 52 (1) (2003) 2–9.

[118] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP)—Round XIII, Proteins: Struct, Funct, Bioinform 87 (12) (2019) 1011–1020.

[119] C. Wu, R.C. Gudivada, B.J. Aronow, A.G. Jegga, Computational drug repositioning through heterogeneous network clustering, BMC Syst Biol 7 (S5) (2013) S6.

[120] H. Huang, T. Nguyen, S. Ibrahim, S. Shantharam, Z. Yue, J.Y. Chen, DMAP: a connectivity map database to enable identification of novel drug repositioning candidates, BMC Bioinform 16 (2015) S4.

[121] H.T. Yang, J.H. Ju, Y.T. Wong, I. Shmulevich, J.H. Chiang, Literature-based discovery of new candidates for drug repurposing, Brief Bioinform 18 (3) (2017) 488–497.

[122] A.P. Chiang, A.J. Butte, Systematic evaluation of drug–disease relationships to identify leads for novel drug uses, Clin Pharmacol Ther 86 (5) (2009) 507–510.

[123] E. Karatzas, G. Kolios, G.M. Spyrou, An application of computational drug repurposing based on transcriptomic signatures, Methods Mol Biol 1903 (2019) 149–177.

[124] H. Xu, M.C. Aldrich, Q. Chen, H. Liu, N.B. Peterson, Q. Dai, et al., Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality, JAMA 22 (1) (2015) 179–191.

[125] R.A. Hodos, B.A. Kidd, K. Shameer, B.P. Readhead, J.T. Dudley, In silico methods for drug repurposing and pharmacology, Wiley Interdiscip Rev Syst Biol Med 8 (3) (2016) 186–210.

[126] F. Cheng, R.J. Desai, D.E. Handy, R. Wang, S. Schneeweiss, A.L. Barabasi, et al., Network-based approach to prediction and population-based validation of in silico drug repurposing, Nat Commun 9 (1) (2018) 1–12.

[127] A.G. Singal, P.D. Higgins, A.K. Waljee, A primer on effectiveness and efficacy trials, Clin Transl Gastroenterol 5 (1) (2014) e45.

[128] B. Munos, Lessons from 60 years of pharmaceutical innovation, Nat Rev Drug Discov 8 (12) (2009) 959–968.

[129] I. Khanna, Drug discovery in pharmaceutical industry: productivity challenges and trends, Drug Discov Today 17 (19–20) (2012) 1088–1102.

[130] Ekins S, Freundlich JS, Coffee M. A common feature pharmacophore for FDA-approved drugs inhibiting the Ebola virus. F1000Research. 2014; 3.

[131] Ekins S, Freundlich JS, Clark AM, Anantpadma M, Davey RA, Madrid P. Machine learning models identify molecules active against the Ebola virus in vitro. F1000Research. 2015; 4.

[132] Z. Zhao, C. Martin, R. Fan, P.E. Bourne, L. Xie, Drug repurposing to target Ebola virus replication and virulence using structural systems pharmacology, BMC Bioinform 17 (1) (2016) 90.

[133] J. Schuler, M.L. Hudson, D. Schwartz, R. Samudrala, A systematic review of computational drug discovery, development, and repurposing for Ebola virus disease treatment, Molecules 22 (10) (2017) 1777.

[134] M. Mottin, J.V.V.B. Borba, R.C. Braga, P.H.M. Torres, M.C. Martini, J.L. Proenca-Modena, C.C. Judice, F.T.M. Costa, S. Ekins, A.L. Perryman, C. Horta Andrade, The A-Z of Zika drug discovery, Drug Discov Today 23 (11) (2018) 1833–1847.

[135] Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, F. Cheng, Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2, Cell Discov 6 (1) (2020) 1–18.

[136] J. Dyall, C.M. Coleman, B.J. Hart, T. Venkataraman, M.R. Holbrook, J. Kindrachuk, et al., Repurposing of clinically developed drugs for treatment of Middle East respiratory syndrome coronavirus infection, Antimicrob Agents Chemother 58 (8) (2014) 4885–4893.

[137] L. Shen, J. Niu, C. Wang, B. Huang, W. Wang, N. Zhu, et al., High-throughput screening and identification of potent broad-spectrum inhibitors of coronaviruses, J Virol 93 (12) (2019) e00023–e119.