Effective holistic characterization of small molecule effects using heterogeneous biological networks

William Mangione[a], **Zackary Falls**[a], **and Ram Samudrala**[a,1]

[a]Department of Biomedical Informatics, University at Buffalo, 77 Goodell st, Buffalo, NY 14203

This manuscript was compiled on April 6, 2022

Drug discovery is the practice of identifying chemical entities with activities useful for treating human diseases. Despite substantial advancements in biotechnologies such as assays for detecting promising hits and computational methods for identifying targets to pursue, the success rate of novel therapeutics in the clinic is rapidly declining. The two most common reasons for drug attrition in clinical trials are efficacy, in that the drug is not capable of treating the disease, and safety, in which the compound is too toxic to patients and does not outweigh any purported benefits. Here we describe the use of an integrated human interactome network to describe the complete picture of drug behavior in biological systems and lead to more promising therapeutic candidate generation. The Computational Analysis of Novel Drug Opportunities (CANDO) platform for shotgun multiscale drug discovery, repurposing, and design was enhanced to include drugs/compounds, proteins, diseases/indications, Gene Ontology entities, and drug side effects, accompanied with various known associations between them, and interactomic signatures were generated for all compounds using the node2vec algorithm for the purpose of relating them in the context of diseases for which they are approved. The results indicate there is significant biological information available in the side effect profile of a drug and they can be used to better relate drugs in terms of their therapeutic behavior. Side effects were also predicted for all compounds in the platform; known drug-side effect associations were recovered at rates ten times higher than what can be expected with random chance. The network architecture was also investigated in the context of protein pathways, where drug distance to various pathways served as the features for input to a random forest machine learning model in the context of indications they are associated to treat, with interesting results highlighted for mental disorders and cancer metastasis. This pipeline highlights the ability of CANDO to accurately relate drugs in a multitarget interactomic context, and paves the way for predicting novel putative drug candidates using the information gleaned from not only their side effect profiles, but their impacts on protein pathways as well.

Bioinformatics | Systems biology | drug discovery | multitargeting | computational drug repurposing

DRAFT

Current drug discovery pipelines are largely ineffective at bringing novel therapeutics to market, on average requiring over a dozen years and more than 2 billion dollars (1). The two dominant reasons for drug attrition in clinical trials are efficacy and severe adverse events, contributing to 56% and 28% of failures, respectively, resulting in a roughly 10-13% success rate (2–4). One reason why so many drugs fail during trials is due to traditional methods often employing the 'magic bullet' approach, where one biological macromolecule (typically a protein) is targeted by a drug with the hypothesis being that modulating its activity will be therapeutic for a patient suffering from the disease (5). A shining example using this rational drug design approach is imatinib, which was developed to target a specific mutated protein in chronic myeloid leukemia and has produced amazing results (6). Yet even this drug has since been discovered to have many other protein targets, leading to a wider range of therapeutic uses (7–10). This "reductionist" view moves away from the functional considerations of these individual protein targets on the whole system and partly explains why many drugs fail due to efficacy (11, 12). On the other hand, adverse events are simply unintended effects of drugs, commonly called 'side effects', and are evidence of the promiscuous nature of small molecule therapeutics. These ulterior effects are often neglected during drug discovery processes, which helps to explain why adverse events are the second leading cause of clinical trial failure.

Many groups have already produced systems biology approaches that aim to integrate greater levels of complexity beyond

Significance Statement

Our study outlines the capability of our drug discovery, repurposing, and design platform known as CANDO to more accurately describe small molecule therapeutic behavior in biological systems via the use of heterogeneous networks. This includes the ability to predict indications as well as side effects for compounds without clinical data, ultimately advancing the science of drug discovery.

W.M. conceived the prediction pipelines, research design, approach and methods, conducted all experiments and analysis, implemented all pipelines, and drafted the manuscript. Z.F. helped with data generation, research design, approach, and methods, and editing the manuscript. R.S. conceived the prediction pipelines, research design, approach and methods, supervised the overall project, and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

The authors declare no competing interests.

[1]To whom correspondence should be addressed. E-mail: ram@compbio.org

the scope of individual proteins with many successes in the area of drug discovery (13–19). In addition, others have employed computational methods to determine associations between proteins, pathways, and structural elements of compounds to adverse drug reactions to better understand the entirety of pharmacological effects drugs have on biological systems (20–22).

Herein is described the use of the Computational Analysis of Novel Drug Opportunities (CANDO) platform for both drug-ADR as well as drug-indication prediction. CANDO is a shotgun multiscale drug repurposing, discovery, and design platform that assesses the therapeutic potential of small molecule compounds in every indication/disease based on their simulated interactions with various proteomes (23–32). The hypothesis is that drugs with similar proteomic interaction "signatures" will have similar effects in biological systems and therefore, can treat the same diseases. The current version (v2) of the platform features thousands of both human-approved drugs and investigational compounds and the diseases for which they are indicated/associated, as well as tens of thousands of protein structures from multiple organisms, such as *Homo sapiens* and SARS-CoV-2. The platform is benchmarked using a protocol that determines how often two drugs indicated for the same disease are considered behaviorally similar based on their proteomic interaction signatures. Aside from varying the exact composition of proteins in the interaction signatures, chemical fingerprints, or vectors tallying the presence or absence of substructures on a molecule, are to date the only other form of data used to characterize similarity between compounds in CANDO. Given the substantial heterogenous data available in biomedical databases regarding biological systems and the phenotypic consequences of modulating the activity of their components, CANDO is well positioned to further elucidate these relationships, especially for small molecule therapeutics and their impacts on proteins, pathways, and diseases. In this study, a biological network was constructed using the human proteome and its known/predicted interactions with a library of 12,951 drugs/compounds as the basis. The network was further supplemented with known interactions between the proteins and their functions in various pathways, as well as their associations to human diseases and entities in the Gene Ontology (GO (33–35)). A comprehensive library of adverse drug reactions (ADRs) extracted from both drug labels and adverse event reports was also included serving as a feature describing, as well as primary outcome of, drug behavior. Topological signatures generated using the node2vec algorithm (36) in various permutations of the network highlighted the rich information provided by the ADR associations for assessing compound similarity in the context of disease, which were then used to predict ADRs for those with unknown profiles. Further, pathways strongly implicated to contribute to mental disorders and metastatic cancers were elucidated via random forest machine learning models, with many of the most important pathways known to be associated to both of the respective diseases, and in the latter case implied that antineoplastic drugs may be simultaneously impacting multiple pathways to enact their therapeutic mechanisms.

Methods

An overview of the study design is outlined in Figure 1. Briefly, multiple heterogeneous databases were integrated into a unified network architecture, which was probed to generate interactomic signatures for drug-indication benchmarking and prediction of drug-ADR associations, as well as gleaning insight into potential pathways responsible for drug therapeutic mechanisms.

Heterogeneous biological data collection.

Drug/compound structure library curation. Drug and drug-like small molecule structures were extracted from DrugBank (37) comprising 13,194 compounds, including 2,449 approved drugs and 2,519 small molecule metabolites, with the remaining 8,226 being either experimental, investigational, illicit, withdrawn, or veterinarian approved only compounds. Biologic therapeutics were excluded from the library. Additionally, a collection of 337 compounds were removed following the protein-compound interaction scoring protocol (see below) due to the lack of variance in their scores caused by their simple chemical composition (elemental ions, inorganic salts, etc), and were deemed uninteresting therapeutic candidates. The final libraries used for network analyses include either this reduced 12,951 compound library, after the addition of 24 compounds of interest implicated in indications for which CANDO has been applied, including 4 GTPase KRAS inhibitors for non-small cell lung carcinoma, or the 2,336 approved drug sublibrary.

Protein structure library curation. Protein structures for the human proteome were extracted from the Protein Data Bank (PDB) (38, 39), if available, or modeled from sequence using I-TASSER (40–42). Among the 19,582 human protein sequences extracted from UniProt (43), solved structures were chosen by 1) identifying all PDB entries that map to the UniProt ID of interest using the SIFTs project (44, 45), 2) clustering them by the indices of their starting and ending sequence coordinates (relative to the full sequence) using the DBSCAN algorithm, and 3) choosing the lowest resolution structures from each cluster that covered the largest portion of the full sequence while minimizing overlap (<50%). Further, solved chains were excluded if they were fewer than 100 residues, but if the full sequence was fewer than 150 residues, this threshold was reduced to 30. This resulted in 4,966 proteins with at least one available solved structure in the PDB, comprising 5,316 total chains. The remaining sequences were aligned to a library of PDB sequences with no greater than 70% sequence identity using the HHBLITS suite (46). Structures were modeled for sequences that matched the following criteria: 1) sequence identity $\leq$ 30%, and 2) number of aligned residues $\geq$ 100 if the sequence length was $\geq$ 150 residues, or aligned residues $\geq$ 30 otherwise. HHBLITS was run with 1 iteration. This resulted in 3,069 modeled structures for a total of 8,385 proteins comprising the human proteome.

Indication/disease library curation. Disease indications were associated to approved drugs using the Comparative Toxicogenomics Database (CTD) (47), comprising 22,772 associations among 2,257 indications. The indication library was further supplemented using the DisGeNET database (48), which extracts protein-disease associations from existing databases as well as the literature

DRAFT

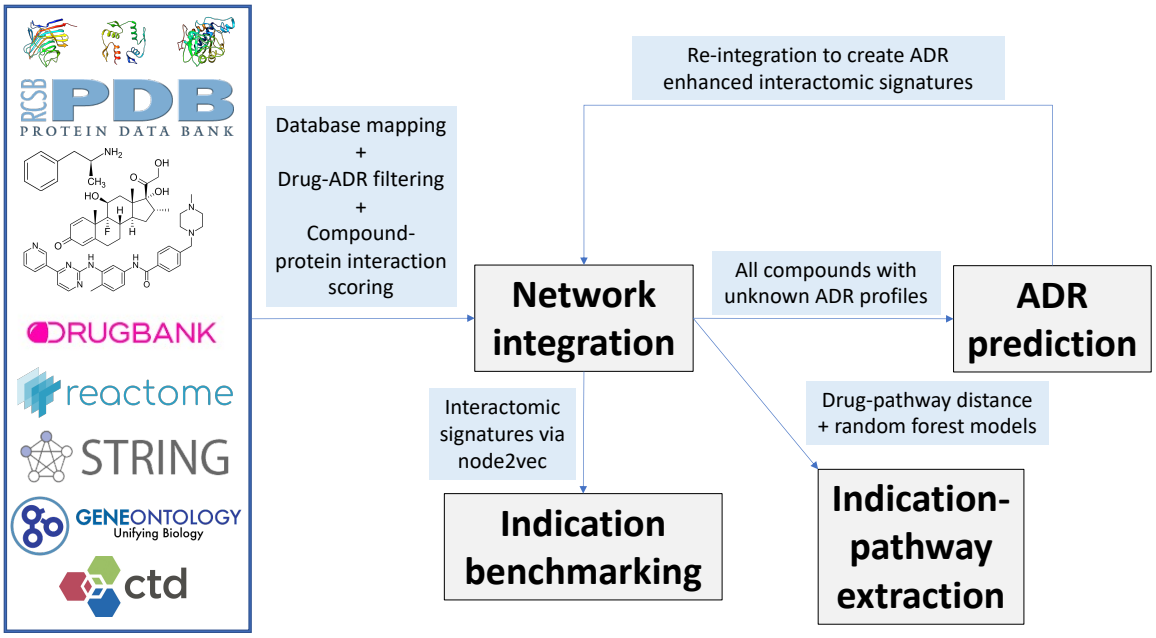

Fig. 1. Overview of the CANDO interactomic signature pipeline and analyses. Several biomedical databases were integrated into a unified network architecture after matching the identifiers in these databases to primarily drugs/compounds from DrugBank and proteins from the PDB and/or UniProt. Prior to integration, filtering of drug-ADR associations from the OFFSIDES database was necessary due to redundancy with the indication mapping from the Comparative Toxicogenomics Database, and a compound-protein intertaction scoring normalization scheme was also devised to supplement the network with additional interactions (see Methods). The interatomic signatures generated from the node2vec algorithm of various permutations of the network were benchmarked to assess how well they relate drugs in the context of disease, as well as to predict ADR associations for compounds with unknown profiles. Finally, drug distance to pathways served as features for input into random forest machine learning models to gain insight into which of them are seemingly important for drug therapeutic mechanisms, which coupled with the results of the indication benchmarking proved these networks are superior to traditional CANDO pipelines for assessing drug-drug similarity/behavior in biological systems.

using advanced text-mining methods. This added 4,961 indications and culminated in 337,177 disease-protein associations, which were split into four subsets according to evidence score thresholds of 0.0, 0.02, 0.1, 0.3, and 0.6. The evidence score is a culmination of the presence of the association in databases or the literature, as well as the consensus among these sources, with higher scores indicating greater confidence. Both databases utilize Medical Subject Headings (MeSH) terms (49), with the total after combining the two data sources reaching 7,217 indications.

Adverse drug reaction library curation. Adverse drug reactions (ADRs), or side effects, of drugs in the CANDO library were extracted from two main sources: 1) the SIDER database (50) which directly annotates the ADRs listed on the label of each drug, and 2) the OFFSIDES project (51), which uses the FDA Adverse Event Reporting System and Canada's MedEffect resource to identify ADRs commonly reported for drugs that are not explicitly listed on drug labels. OFFSIDES uses statistical techniques to identify additional ADRs for a drug that are not simply signs or symptoms of a disease for which a given drug is commonly indicated.

Of the 1,430 drugs present in SIDER, 686 were mapped to our compound library for a total of 92,143 associations to 5,144 ADRs. Due to the source of these associations being directly from drug labels, these associations were treated as a gold standard and were not further filtered.

Two separate versions of the OFFSIDES database were mapped to the CANDO drug library, with the first version sourcing ADR reports up to 2009, and the second version extending to 2014. After synthesizing these sets by matching the MedDRA ADR identifiers (52) of the second release to the Unified Medical Language System (UMLS) ADR identifiers (53) of the first release, 1,622 drugs in CANDO were associated to 10,268 unique ADRs for a total of 191,591 associations.

Further filtering of this set was necessary due to the significant number of drugs associated to ADR terms that mapped to identical SNOMED concepts as those for the diseases for which they are indicated to treat. To prevent biasing the network by including these drug-ADR associations that could artificially enhance drug-indication prediction performance, all ADRs and indications were first mapped to their equivalent SNOMED Clinical Terminology (54) concepts using the UMLS Metathesaurus, and only ADR concepts of the "disorder" class were kept. Drug-ADR associations were removed if the mapped SNOMED concept for the ADR term directly matched the mapped SNOMED concept for the MeSH indication for which the drug is indicated. Further, for a given indication I and its associated drug set D, all of the indications associated to the drugs in D were grouped into a set X; if any of the mapped SNOMED concepts for the indications in X directly matched the mapped SNOMED concept for an ADR, associations to this ADR for all drugs in D were removed (Figure 2A). This was repeated for all indications in CANDO. Lastly, the concept hierarchy from SNOMED was used to perform a filtering process that utilized "is_a" relations to connect the lower level ADR concepts to their upper level classes (e.g. both "Low blood pressure" and

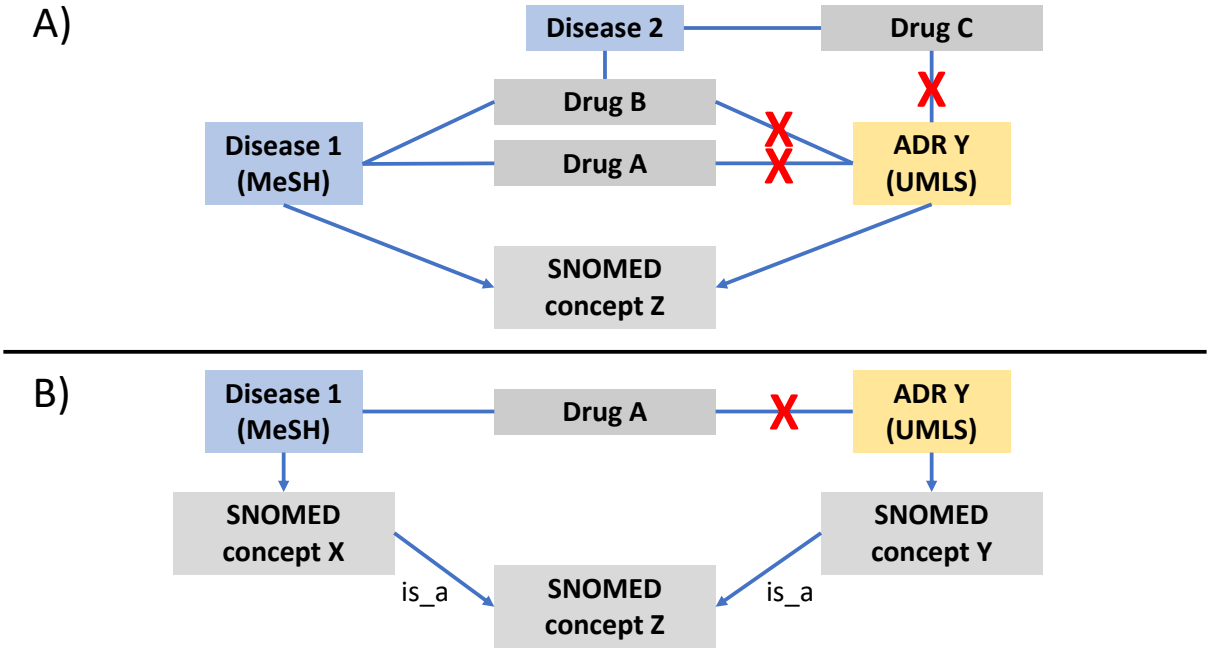

Fig. 2. Schematic representation of removed ADR associations in the network. The drug associations to various ADRs were removed (red Xs) if they matched two different criteria. A) If the disease term from the CTD (annotated in MeSH) mapped to the same SNOMED concept as the ADR term from Offsides (annotated in UMLS), the association to that ADR was removed, signified by both Drug A and Drug B. But even further, this was repeated for any drug that shared a disease association with any drug from the originally matched disease, signified by Drug C. B) If the ADR and disease terms mapped to their respective SNOMED concepts and both of those mapped to the same upper-level SNOMED concept, using the is_a relations in the SNOMED hierarchy, the association was removed. This scheme was used to eliminate as many false positives as possible which in the network would be nonsensical relations.

DRAFT

"Heart disease" belong to "Disorder of cardiovascular system"). Similarly to the previous process, sets were created by grouping all mapped SNOMED concepts of the ADRs and indications that belonged to the same upper level SNOMED concept; if a drug was associated to both an indication and ADR of the same upper level class (excluding top level terms "Disease" and "Clinical finding"), the ADR association was removed (Figure 2B). This culminated in 50,210 associations between 1,112 drugs and 1,413 ADRs, after removing all ADRs with fewer than 10 drugs associated. This was combined with the 85,095 associations from SIDER that could be mapped to SNOMED for a final, non-redundant set of 133,322 associations between 4,245 ADRs and 1,171 compounds.

Protein-protein and protein-pathway associations. All human proteins were associated to their known interactions with other proteins and the biological pathways to which they belong using the STRING and Reactome databases (55, 56), respectively, including those without a solved or modeled structure available. STRING collates protein-protein interactions (PPIs) from multiple sources, including the literature and other databases, evolutionary information transferred from known PPIs in other organisms, computational predictions, and various others. A cumulative score is assessed based on the available evidence for a protein pair; in this study, a cutoff of 700 was used to determine a PPI, which is deemed high confidence according to the curators, resulting in 169,208 interactions. Reactome constituted 2,219 pathways comprising 108,466 associations, with an average of 48.8 structures per pathway.

Gene Ontology annotations. As an additional resource, annotations to the Gene Ontology (GO) for all human proteins as provided by the UniProt database were extracted for all descendant terms of the three main upper level entities ("biological process", "molecular function", and "cellular component"). This resulted in 18,094 unique GO entities and comprised 262,059 total annotations.

Compound-protein interaction scoring protocol. Interaction scores between each protein and compound structure were computed using an in-house bioanalytic docking protocol known as BANDOCK. This protocol relies on two main components to assign a real-valued score between a compound and protein: 1) known or predicted binding sites on the protein, as well as the ligands known to bind to these binding sites as determined by X-ray structures in the PDB, and 2) chemical fingerprinting methods that identify the presence or absence of molecular substructures on a compound. The first component is handled by the COACH algorithm of the I-TASSER suite, which utilizes a consensus scoring approach of three different methods to score how similar the binding sites on the query protein are to the binding sites in the PDB. The output of COACH is a list of binding sites, the ligands known to bind to them, and an associated score (Pscore) for each. To assign a score for a compound-protein pair, the Sorenson-Dice coefficient between the ECFP4 chemical fingerprints of each ligand associated to the predicted binding sites for the protein and the query compound (Cscore) are computed; the strongest matching ligand

(Cscore closest to 1.0) is chosen and multiplied by the Pscore for the binding site to which this ligand is associated (Pscore x Cscore). This process is iterated for all compound and protein pairs to generate proteomic interaction signatures for all drugs/compounds in the library.

Benchmarking protocol. The main goal of the CANDO platform is to accurately describe drug/compound behavior, and the primary protocol used to assess the performance of a particular pipeline utilizes the indication approvals extracted from the CTD to determine how often drugs associated to the same indication are deemed 'similar' according to various thresholds/cutoffs. In the traditional pipeline, similarity of two compounds is calculated by comparing the scores in their proteomic interaction signatures using the root-mean-square-deviation (RMSD) metric. This is repeated iteratively for all pairs in the library, resulting in each having a list of compounds ranked by decreasing similarity. For each drug approved for a given indication, the benchmarking protocol averages how many times another drug associated to that indication is found within a certain rank cutoff (10, 25, 100, etc) in the list of most similar compounds to the hold-out drug; a topK accuracy is assessed for the indication based on the number of times another associated drug was found in the top K most similar compounds for each drug. This is repeated for all indications to determine the global topK accuracy, and then repeated for all other cutoffs. In addition to the average indication accuracy (AIA) just described, the average pairwise accuracy (APA) and the indication coverage (IC) are computed, which are the weighted average based on how many drugs are associated to each indication and the number of indications with non-zero accuracies, respectively. This protocol resembles the "hits@K" metric but differs in that it does not consider the total number of hits in the top K for a single drug-indication pair, but rather the presence of *at least one* within the top K. For this study, the benchmarking protocol only included the sublibrary of 2,336 approved drugs and used cosine distance as opposed to RMSD (for substantial increase in computational speed).

Compound-protein interaction selection. Despite having >100M interaction scores between every protein and compound in CANDO, including them all would be meaningless due to the unweighted, binary nature of the edges in the network architecture. Further, smaller compounds tend to be favored by the ECFP4 fingerprint, while many others structurally resemble ligands commonly co-crystallized with proteins in the PDB, both of which inflate their Cscores. In addition, well studied proteins like cancer targets or kinase families also tend to be overrepresented in the PDB and therefore have greater Pscores from COACH, as well as increased ligand diversity as a consequence. Since the Pscore from COACH implicitly measures confidence in the binding site and therefore its associated ligand, simply considering only the Cscore (as opposed to multiplying it by the Pscore) during the BANDOCK scoring protocol would 1) disregard a significant portion of the embedded evolutionary information and 2) reduce the discernability between a compound with the same matching ligand for multiple proteins. Therefore, a normalization scheme was derived to select the strongest and most unique interactions between all compounds and proteins. For a given interaction score between a compound and protein, the min-max normalized score was computed based on the non-zero minimum and maximum values among all interaction scores for the corresponding protein; this process was iterated for all proteins in the compound signature. The normalized scores were then ranked for the compound and the top 5 interactions were initially selected to be included in the network. If a tie occurred at ranks 5 and 6, all subsequent interactions equal to the value at rank 5 were also selected. This process was repeated for every compound in the library. To further reduce any bias that may still be present even after normalization, all proteins were reduced to their top 50 compounds; if ties occurred at rank 50, all tying compounds were ranked by the inverse of their average interaction score against all proteins, which therefore prioritized more unique interactions. This culminated in 50,345 interactions across 6,161 (73.5%) protein structures.

Compound-protein interactions annotated in DrugBank were also included in the network, constituting 13,702 interactions, of which 1,876 (13.7%) were captured in the normalization scheme. This resulted in 62,510 unique compound-protein associations after combining the interactions from multiple chains of the same protein.

Network integration and preliminary analyses. The following biological entities and relations were initially included in the network for preliminary analysis: drugs/compounds, proteins, indications, pathways, GO terms, ADRs, PPIs, compound-protein interactions, compound-ADR associations, protein-pathway associations, and protein-GO annotations (Figure 3). Since this study aimed to improve drug repurposing benchmarking accuracies, drug-indication approvals were excluded from all networks. The complete network featured 64,308 nodes and 1,072,280 edges and was built using the Python package NetworkX. Separate networks were built for each permutation of including or excluding the relation types above, except for the protein-indication associations in which the DisGeNET thresholds were varied instead, totaling 80 unique networks.

For each network architecture, node representations/embeddings were generated using the Python implementation of the node2vec algorithm, which uses biased random walks starting from each node to produce a continuous value vector in a lower dimensional space that aims to preserve the contextual neighborhood in which the node is situated. The walk length, number of walks, and output dimensionality were set to 10, 100, and 128, respectively, for all networks. To assess how well the node embeddings captured the therapeutic potential of compounds in the network, the vectors for the approved drugs were used as input to the benchmarking protocol in place of their traditional protein interaction signatures. Drug-drug similarities and drug repurposing accuracies were computed as described in the benchmarking protocol section (above).

Compound-ADR prediction protocol. ADRs were predicted for all 12,951 compounds in the CANDO library based on the similarity of each compound to drugs with known ADRs. Signature similarity is first computed for all compound-compound pairs in the library in the same way as for the benchmarking protocol. The ADRs associated to the most similar compounds within a set cutoff X to a query compound are counted, and a probability is assessed based on the number of times the

DRAFT

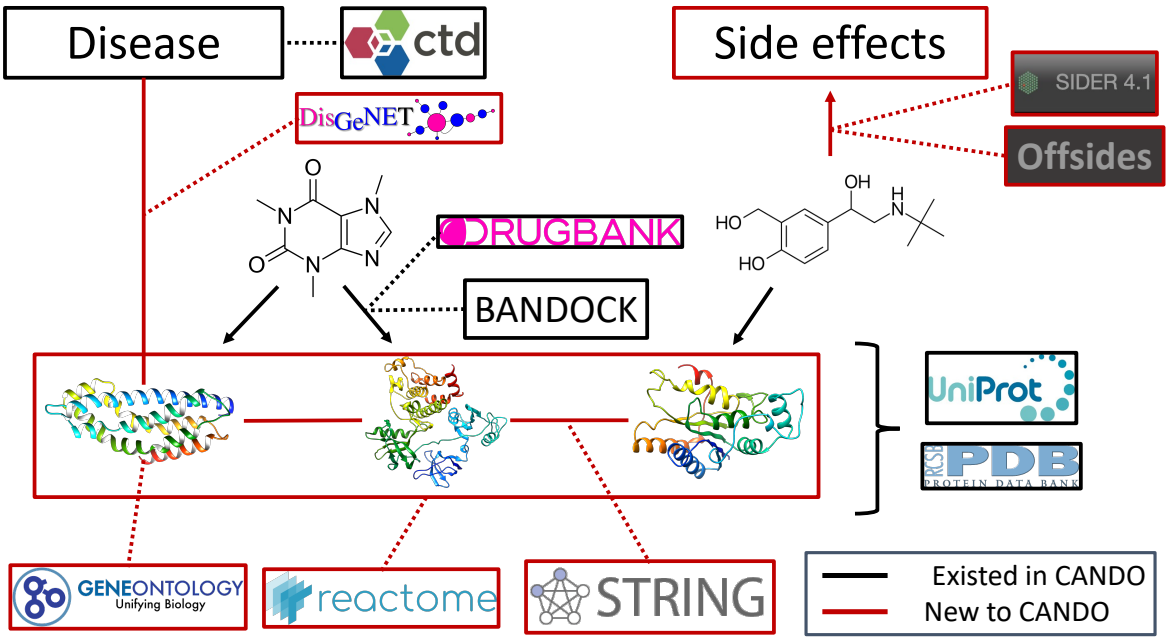

Fig. 3. Overview of architecture and data sources for heterogeneous network. Proteins and their structures were extracted from UniProt and the Protein Data Bank, which were connected to each other via protein-protein interactions from the String database as well as to protein pathways from the Reactome database. Protein annotations to Gene Ontology terms were extracted from UniProt as well. Proteins were mapped to diseases using the DisGeNET database, with additional diseases being sourced from the Comparative Toxicogenomics Database. Compound and drug structures were downloaded from DrugBank in addition to known protein-drug interactions, which was further supplemented using our BANDOCK interaction scoring protocol. Finally, side effects, or adverse drug reactions, for drugs were extracted from the SIDER and Offsides databases. Everything in black was an already-existing entity or relation in CANDO, whereas everything in red signifies a novel association for this network-based analyses and constitutes most of the connections between these entities.

ADR was present, the frequency of the ADR within the library, the size of the compound library, and the cutoff X using the hypergeometric distribution. The ADRs are ranked by the probability, which represents the chance of randomly selecting at least that many compounds associated to the ADR without replacement in X trials.

Pathway importance via random forest models. The importance of each pathway in predicting drug-indication associations was assessed via random forest machine learning models. For 60 indications from the CTD mapping with at least 50 drugs associated, a random forest model was trained using 90% of the associated drugs as the positive samples, the remaining 10% as the positive test samples, and an equal number of training and test "negative" samples chosen at random from the sublibrary of drugs with at least one indication association, ensuring that the randomly chosen drug is not already associated to the indication or any upper level class to which the indication belong. The features of the drugs were a binary vector of integers with a 1 meaning the distance of the drug to the pathway in the network is less than or equal to 3, otherwise 0. The pathways were filtered for those that satisfied the following criteria: 1) at least 5 and no greater than 250 proteins associated, 2) at least two associated proteins with at least one interaction to any compound in the network, and 3) at least 1% and no greater than 33% of compounds with a distance of 2 to the pathway. Ten iterations were performed for each indication; the average performance over all iterations was assessed and the feature importances for the indication models with the best average performance were extracted.

Results and discussion

Preliminary benchmarking accuracies for network permutations. All permutations of the network, by varying the DisGeNET protein-indication score threshold and either including or excluding all other relations, were benchmarked by replacing the traditional proteomic interaction signatures with the 128 dimension embeddings for the 2,336 approved drugs generated via node2vec, herein called "interactomic signatures". A significant disparity in performance was observed for signatures derived from networks that included adverse drug reactions (ADRs) with the mean top10 average indication accuracy (AIA; 30.0%) exceeding that of the signatures from networks excluding ADRs of 21.8% (37.3% of %). No significant differences were observed in other permutations of the network outside of the ADRs. However, all interactomic signatures outperformed the control proteomic interaction signature matrix, which achieved a top10 AIA of 20.7%.

To assess the relative impacts of other relations on benchmarking accuracy that may have been dampened by the significant contribution of the ADR associations, a new compound sublibrary was created containing only drugs with at least one ADR and indication association. After benchmarking with only this reduced approved sublibrary of 1,047 drugs, the interactomic signatures from networks including ADRs still far exceeded the performance of those that did not, though not as dramatically,

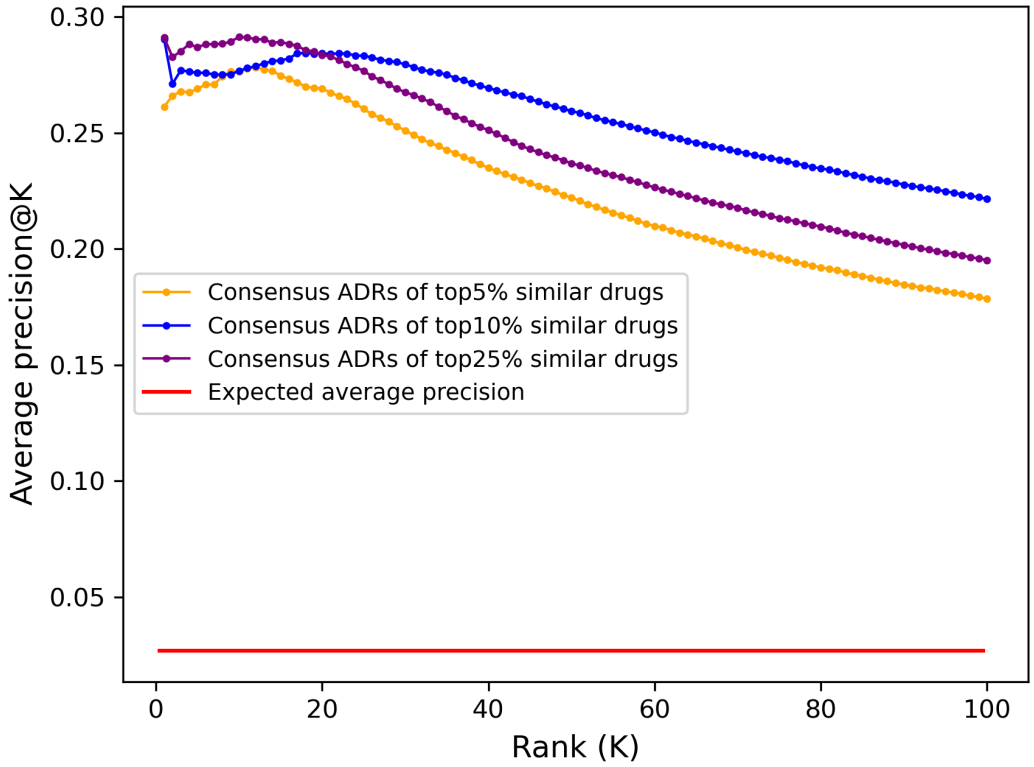

Fig. 4. Drug-ADR association precision for known ADRs predicted from CANDO interactomic signatures. The top 100 adverse drug reaction (ADRs) predictions for all 12,951 compounds in the CANDO library were generated using a consensus voting scheme. The ADRs associated to the most similar compounds, based on the similarity of their interactomic signatures generated using node2vec in a network devoid of all ADR associations, within a determined cutoff to a query compound were counted; the predictions are ranked by the probability computed using the hypergeometric distribution based on how many times the ADR was associated to the compounds most similar to the query, the number of ADRs present in the platform, and how many most similar compounds were inspected. Average precision was calculated at a given rank K by averaging the precision@K for all drugs with known ADRs (1171). The predictions generated using a cutoff of top 5% (647; orange) have lower average precision across all ranks compared to those generated using a cutoff of top 10% (1295; blue), yet both are 10-fold more precise on average than what is expected at random (red). The predictions using a cutoff of 25% (3238; purple) had higher precision at early ranks, peaking at rank 10, but significantly decreased at progressive ranks. Similarly, aside from rank 1, the average precision of the top 10% lists peaks at rank 18 and steadily declines thereafter, while the top 5% lists peak earlier at rank 12; this is likely due to the scarcity of drugs with known ADR associations in the full library. The dramatic increase on the random expected average precision for all sets of predictions indicates the interactomic signatures capture compound behavior effectively.

with a mean top10 AIA of 33.8% versus 27.6% (22.5% of %). In addition, a slight drop in performance was observed for the signatures that included protein-GO annotations and used DisGeNET thresholds of 0.0 and 0.02, so for subsequent analyses the GO annotations were excluded and the 0.1 threshold was utilized as it provides the greatest coverage of protein associations for indications used for benchmarking while still maintaining competitive accuracies. In addition, all network signatures exceeded the performance of the control proteomic interaction signature matrix, which achieved a top10 AIA of 25.7% for this sublibrary.

Compound-ADR prediction and complete interactomic signature benchmarking. Due to the strong biological signal provided by the ADRs in the context of drug repurposing, ADRs were predicted for all compounds in CANDO using a consensus voting scheme (see Methods). The top 100 ADRs identified by tallying the most overrepresented ADRs in the associations of the top 5, 10, and 25% (647, 1,295, 3238, respectively) most similar drugs to a query compound were generated. To determine the most appropriate cutoff for selecting how many ADRs were to be associated to the compounds in the network, the precision@K was computed for all ranks between 1 and 100 and averaged at each rank for all drugs with existing ADR associations from OFFSIDES and SIDER 4. Aside from rank 1, which is the maximum average precision for the top 5 and 10% cutoffs, the peak average PR@K was at rank 10 for the top25% cutoff; however, due to a bias discovered in the drug-indication mapping with disproportionate impact on benchmarking accuracy (see below), for further analyses we selected the top10% cutoff due to its better performance at higher numerical ranks. Its peak performance was at rank 18; rounded this up to the top 25 ADRs due to comparable performance and associated these top ADRs to all compounds in CANDO with no ADR associations. For drugs with existing ADRs that totaled fewer than 25, the ADRs were iterated in rank order until 25 unique ADRs were associated to the drug. In addition to this top25 set, another set of all 100 ADR predictions were extracted in the same way for comparison.

Embeddings were generated with node2vec again using the complete compound library, their newly associated top25 or top100 ADRs, PPIs, protein-pathway associations, and a DisGeNET score threshold of 0.1. The embeddings for the 2,336 drugs

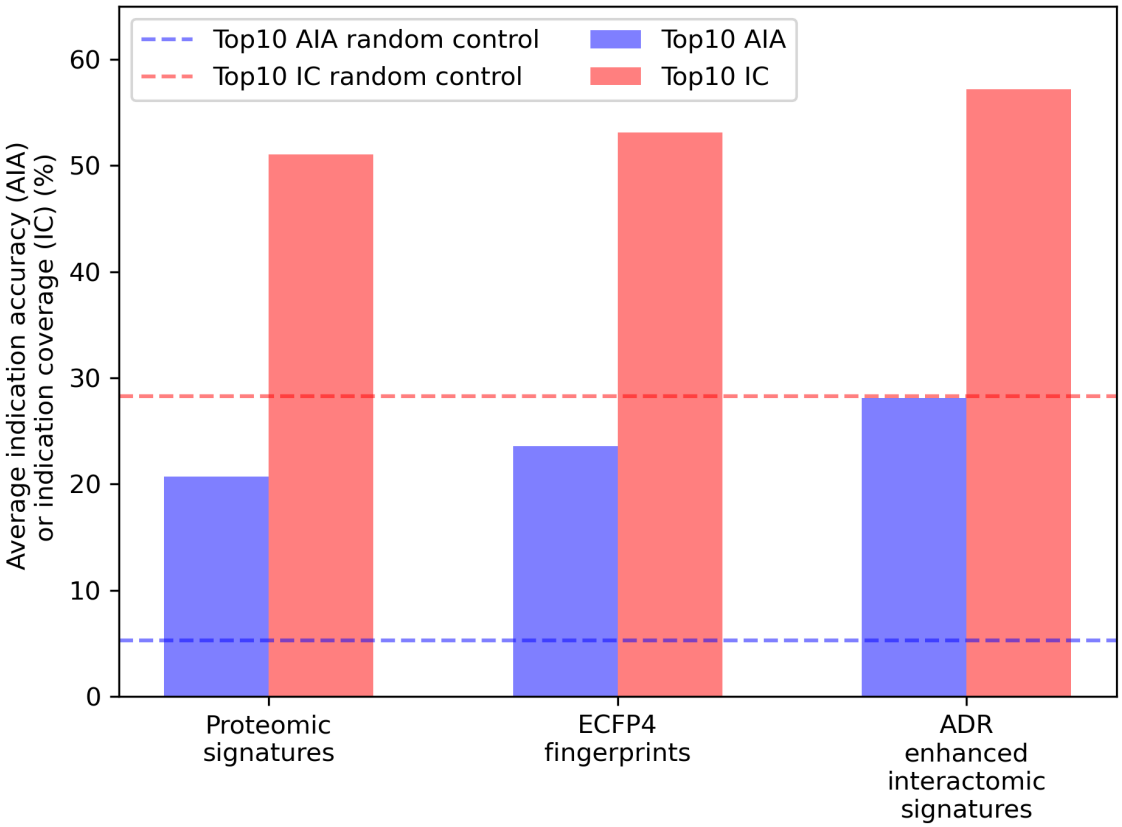

Fig. 5. Benchmarking performance for various CANDO pipelines. The ability of different CANDO pipelines to accurately determine similarity of drugs known to treat the same indications was assessed using a hold-one-out benchmarking scheme on a per indication basis. All indication accuracies are averaged to compute the average indication accuracy (AIA; blue), while the number of indications with a non-zero accuracy, or indication coverage, is also counted (IC; red). The performance of the proteomic signature pipeline using a library of 8,385 human proteins and the ECFP4 molecular fingerprint pipeline are depicted and compared to the human interactome network pipeline, which uses the node2vec algorithm to output continuous value low dimensional embeddings based on connectivity of the heterogeneous biological network. The network pipeline also featured a set of the top 25 predicted drug-ADR associations for all compounds generated using a consensus voting scheme (see Methods). The AIA of this interactomic pipeline of 28.1% exceeds those of the proteomic signature and ECFP4 fingerprint pipelines of 20.7% and 23.6%, respectively, and is over five times that of the random control value of 5.3% (dashed blue). The same is true for IC where the ADR enhanced interactomic signatures achieve 57.2%, amounting to 97 additional indications for which there is a non-zero accuracy (out of 1,588), with a random control expected value of 28.3% (dashed red). This further indicates the interactomic signatures are capturing compound behavior more effectively than the less dynamic proteomic signature and fingerprint pipelines, and demonstrates that considering compound-protein interactions in the their complete biological context is beneficial for understanding their therapeutic mechanisms.

in the approved sublibrary were extracted and subjected to the benchmarking protocol as before. The top25 ADR network achieves a top10 AIA of 28.1%, exceeding that of the top100 ADR network of 25.8%; both of them exceed the aforementioned proteomic signature control performance of 20.7%. Both ADR networks also outperformed another control matrix that uses only the Dice coefficients between drugs in the library as the compound-compound similarity scores, which achieves a top10 AIA of 23.6% for the approved drug sublibrary. These fingerprint only pipelines typically outperform all other proteomic based pipelines in benchmarking, further indicating the interactomic signatures are capturing significantly more therapeutic signal.

In addition to AIA, both the average pairwise accuracy (APA) and indication coverage (IC) dramatically increase for the ADR enhanced networks relative to control. Top10 APA is the percent of successful hits within the top 10 among all drug-indication pairs in the platform; the top25 ADR enhanced interactomic signatures achieved a top10 APA of 46.7% compared to the control proteomic signature and fingerprint pipeline values of 35.7% and 38.7%, respectively, which constituted a raw increase in 2,070 and 1,502 successful top10 drug-indication hits, respectively. The number of indications belonging to each upper level MeSH class for which the network pipeline achieved increased benchmarking accuracies relative to each control pipeline were counted, and a probability (p) was assessed based on the number of indications used for benchmarking (1588), the frequency of each upper level class among those indications, the number of indications belonging to each class with improved scores, and the total number of improved indications using the hypergeometric distribution. This was also repeated for the indications with lower accuracies relative to the controls. Significance was determined by a probability $\leq$0.005. A total of 616 indication accuracies were improved relative to the proteomic signature pipeline and 550 were improved relative to the fingerprint pipeline 6. The strongest improvement among both controls was in the "Neoplasms" ('C04') class, which has 216 indications associated. The interactome pipeline successfully improved benchmarking accuracies for 133 ($p < 1e\text{-}12$) and 125 ($p = 1e\text{-}13$) indications belonging to the C04 class compared to the proteomic signature and fingerprint pipeline, respectively. On the other hand, the greatest decrease in benchmarking accuracy relative to the controls was in the "Infections" ('C01') class

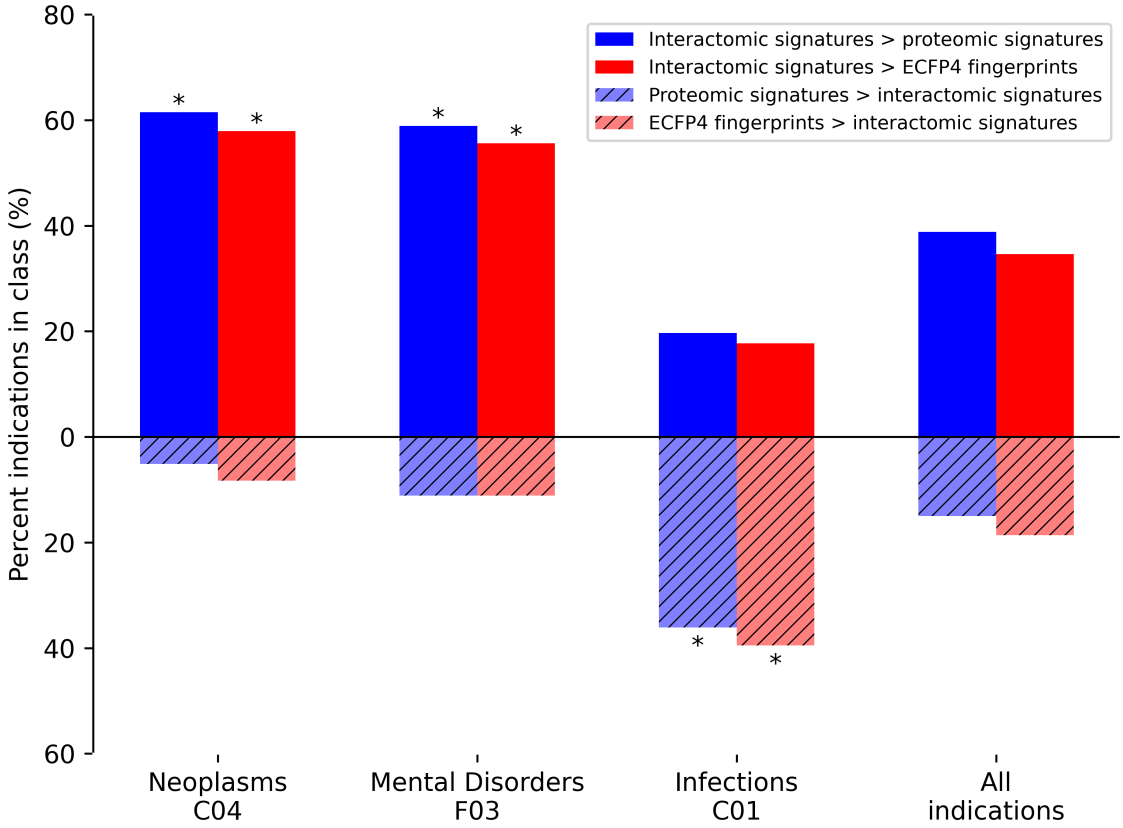

Fig. 6. Comparative benchmarking performance for various CANDO pipelines in specific indication class. The most overrepresented classes of indications with changes in benchmarking accuracy for the interactomic signatures compared to both the ECFP4 fingerprints proteomic signatures are depicted. The average indication accuracy of the interactomic pipeline of 28.1% exceeds those of the proteomic signature and ECFP4 fingerprint pipelines of 20.7% and 23.6%, respectively (see Methods and Figure 5). The interactomic signatures dramatically outperform the proteomic signatures (solid blue) and ECFP4 fingerprints (solid red) in both Neoplasms and Mental Disorders. The proteomic signatures and fingerprints do, however, outperform the interactomic signatures (lined blue and red, respectively) in Infections. Given that drugs treating infectious disease typically target pathogenic proteins, which are not included in the human interactome network, this performance is not surprising. The proteomic signature pipeline achieving significantly better performance, despite being composed of interactions with human proteins, is a phenomenon observed in previous CANDO publications in which the increase in the total number of proteins in an interaction signature leads to predictable increases in accuracy, most likely due to an increase in the coverage of the protein fold space (24). Coupled with the substantial increases in accuracy for complex diseases like cancer and psychiatric disorders, which have various and sometimes poorly understood etiologies, it is apparent that the interactomic signatures are capturing drug behavior beyond just their immediate protein targets, but in the entire context of interactions in which they are situated in biological systems.

to which 147 indications used for benchmarking belong. The proteomic signature and fingerprint pipelines achieved higher accuracies for 53 (p = 3e-12) and 58 (p = 1e-10) of these indications, respectively, compared to the interactomic signatures.

Despite featuring ADR associations for all drugs in the CANDO library, the top25 ADR enhanced interactomic signatures were still well exceeded by the signatures generated using the equivalent network with incomplete compound-ADR coverage in benchmarking accuracy (28.1% vs. 30.6%, respectively). This was inconsistent with previous results indicating that ADRs provide rich biological information, so the introduction of the predicted drug-ADR associations should have increased benchmarking performance for the entire approved drug sublibrary. The exact top10 drug-drug-indication hits responsible for the increase in accuracy were inspected; of the hits produced by the incomplete ADR network that were not produced by the complete ADR network, 74.6% were drug-drug pairs where either both had or both did not have ADR annotations, indicating this pipeline preferentially ranked compounds relative to each other based on the quantity of ADRs associated. Given the split in the approved drug sublibrary is essentially even between drugs with ADRs (1,171) and those without (1,165), this is strong evidence of this network being overtrained, and justified our decision to use a higher ADR prediction cutoff (top25) versus the lower cutoffs with greater precision (≤ 10) to reduce bias (Figure 4).

Drug-pathway association. Random forest machine learning models were trained for all indications with at least 50 drugs associated using the distance to a subset of 821 pathways as features for the drugs. The vectors were binarized with a value of 1 indicating the drug is within a distance of 3 to the pathway, otherwise 0. Each model was trained using a 90/10 split between training and testing samples for ten independent iterations. Among the 60 indications with at least 50 associated drugs, the top 5 with highest accuracy included "Mental Disorders" (MeSH:D001523), "Migraine Disorders" (MeSH:D008881), "Schizophrenia" (MeSH:D012559), "Dyskinesia" (MeSH:D004409), and "Neoplasm Metastasis" (MeSH:D009362) with average accuracies over 10 iterations of 0.83, 0.78, 0.77, 0.77, and 0.76, respectively. The pathway features most important for predicting

Table 1. Pathways most important for predicting Mental Disorders. Pathway importance as extracted from random forest models trained on drugs known to treat Mental Disorders are listed. The models for this indication, which has 54 drugs associated, achieved an average accuracy of 0.83. The ranks of each pathway belonging to each group, their Reactome identifiers, and encompassing description of each group are listed accompanied by literature references supporting their importance in psychological disorders. These pathways, aside from cytochrome P450 substrate metabolism, are well-known targets of many antipsychotic drugs and provides evidence that the pathways extracted by the random forest models are relevant to therapeutic mechanisms.

Ranks	Reactome identifiers	Pathway group description	N Drugs	Evidence
1	HSA-390651	Dopamine receptors	4	(57, 58)
2	HSA-375280	Amine ligand-binding receptors (GPCRs)	14	(59–61)
3	HSA-1296071	Potassium channels	22	(62–64)
4	HSA-390666	Serotonin receptors	5	(58, 65)
5	HSA-211958	Miscellaneous substrates of cytochrome P450s	54	(66)

Table 2. Pathways most important for predicting Neoplasm Metastasis. Pathway importance as extracted from random forest models trained on drugs known to treat Neoplasm Metastasis are listed. The models for this indication, which has 68 drugs associated, achieved an average accuracy of 0.76. The ranks of each pathway belonging to each group, their Reactome identifiers, and encompassing description of each group are listed accompanied by literature references supporting their importance in cancer. All pathways have multiple studies supporting their relevance, suggesting these antineoplastic drugs may be impacting these pathways as part of their therapeutic mechanisms.

Ranks	Reactome identifiers	Pathway group description	N Drugs	Evidence
1	HSA-156581	Methylation via S-adenosylmethionine	63	(67–69)
2; 3; 5; 8	HSA-9018677; HSA-9018678; HSA-9018682; HSA-9027307	Biosynthesis of specialized proresolving mediators (SPMs) and maresins	64	(70–72)
4	HSA-5423646	Aflatoxin activation and detoxification	62	(73–75)
6; 7	HSA-2142670; HSA-2142691	Synthesis of epoxy- and dihydroxy- eicosatrienoic acids, leukotrienes, and eoxins	63	(76–79)
8	HSA-77289	Mitochondrial Fatty Acid Beta-Oxidation	64	(80–82)

these drug-indication pairs were extracted for the top two (distinct) indications, "Mental Disorders" and "Neoplasm Metastasis", with their top 5 non-redundant pathways listed in Table 1 and Table 2, respectively.

Even though the initial benchmarking results using signatures derived from the network with incomplete ADR associations outperform those of using signatures from the complete network, the average number of indication associations for the drugs in the original SIDER and OFFSIDES ADR mapping is nearly an order of magnitude greater than the approved drugs which are absent (16.4 vs. 1.7, respectively). The embedded signatures in the initial incomplete ADR network were likely overtrained and were identifying a correlation between drugs with well characterized side effect profiles and the number of indications for which they are approved, regardless of the exact ADRs to which they were associated; this is evidenced by the relative diminished impact the ADR associations had on benchmarking accuracy using the sublibrary of drugs with at least one ADR association compared to the the equivalent networks without ADRs. The tremendous impact the ADRs have on benchmarking indication accuracies, and therefore describing drug behavior in totality, is likely due to 1) the sheer number of compound "effects" that are added by the ADR associations being nearly six times greater than those available in the indication mapping from the CTD and 2) an increase in granularity for phenotypic outcomes after exposing these compounds to biological systems. The former is an important limitation of most drug-indication prediction methods, i.e. a lack of a significant amount of positive samples from which we can glean information about their mechanisms. The addition of the tens of thousands of ADRs for drugs in the study provided a significant boost in direct phenotypic outcomes. The latter can range anywhere from routine diagnostic findings, such as "Blood urea abnormal", which is likely due to impacts on one or a few protein pathways, to higher level ADRs, such as "stomach cramps" or "headache", which are likely due to multiple contributing pathways. This will be explored in a future work in which the pathways responsible for causing ADRs will be extracted in a similar manner to how they were for the indications in this study, and clustered based on what pathways are seemingly contributing to those outcomes. We believe the interactomic signatures using these ADRs are taking advantage of this granularity and learning compound behavior at multiple levels of complexity, showcased by their enhanced ability to relate drugs approved for the same diseases.

The network pipeline successfully improved benchmarking accuracies for 133 ($p < 1e\text{-}12$) and 125 ($p = 1e\text{-}13$) indications belonging to the C04 class compared to the proteomic signature and fingerprint pipeline, respectively, strongly indicating the contextual biological information of the network more accurately captures oncological therapeutic behavior compared to the linear proteomic signatures or the protein-agnostic fingerprints. This is also true for other diseases with complex etiologies like cardiovascular conditions and mental disorders. However, the proteomic signature and fingerprint pipelines achieved higher accuracies for 53 ($p = 3e\text{-}12$) and 58 ($p = 1e\text{-}10$) of the 147 indications belonging to the "Infections" class, respectively, compared to the network pipeline. This is likely due to the mechanism of most drugs used to treat infectious diseases (antibiotics, antiparasitics, antivirals, etc) being to directly disrupt pathogenic proteins/processes, none of which are included in the human focused biological network in this study. This also provides further evidence that the compound embeddings produced from the network are capturing behavioral signal beyond just the scope of the few immediate proteins with which they interact (as in their linear proteomic signature), and are able to understanding the context of these interactions relative to the human-based diseases

DRAFT

for which they are therapeutic. The proteomic signature pipeline achieving significantly better performance for indications, despite also not containing any pathogenic protein structures in its signatures, is a phenomenon observed in previous CANDO publications in which the increase in the total number of proteins in an interaction signature, regardless of organismal source, leads to predictable increases in accuracy. This is most likely due to an increase in the coverage of the protein fold space and a consequence of evolutionary conservation between protein structure and function (24). This hypothesis will be evaluated in future work where various pathogen enhanced biological networks will be constructed with known protein-protein interactions responsible for their infective mechanisms and probed in the same interactomic manner as the human only network used in this study. The promise of this method extends beyond just direct inhibition of the pathogenic proteins as well; consider a drug such as dexamethasone, which is now a primary treatment for hospitalized COVID-19 patients: the mechanism of this drug is to reduce the severe inflammation cascade (cytokine storm) resulting from viral processes, but is not a direct inhibitor of SARS-CoV-2 proper. A well-constructed network connecting known viral proteins to inflammatory pathways in humans will be able to decipher these relationships and predict effective host-based treatments in addition to direct anti-pathogen inhibition.

Limitations

Despite the significant improvement in benchmarking indication accuracies outlined in this study, there are still many limitations. The reliance on a set of drugs with known ADR associations amounting to less than one-tenth of the total compounds analyzed in this study may be biasing the set of ADRs ultimately being associated to them. Further, a significant amount of information present from the OFFSIDES project had to be removed due to an inability to confidently determine that the drug-ADR associations were not a known symptom/sign of the disease for which the drugs are common treatments. Future work that painstakingly utilizes a biomedical entity hierarchy even more so than was done with SNOMED CT in this study or assessing the co-linearity of ADRs for drugs indicated for the same diseases will garner much more data crucial for understanding compound behavior.

Related to this, the lack of negative samples for both ADRs and indications, especially, likely hinders the performance of the machine learning models used in this study due to an imperfect random sampling of compounds without those labels as the "negative" samples used for training. Curating a library of negative drug-effect associations may be possible for one or a handful of indications, but producing a library on the scale of the mapping provided by the CTD will require extensive automated parsing of the biomedical literature and clinical trial data that is out of the scope of this study.

Based on the architecture of the network, it is not obvious which compound-pathway impacts are that of an agonistic or antagonistic nature. The same is true for the docking protocols used for supplementing the network in addition to the known compound-protein interactions. This is already being addressed in a future work in which the LINCS1000 and Connectivity Map gene expression databases are being integrated to better determine whether a compound is activates a pathway based on the gene expression its members, or vice versa (83, 84). The node2vec algorithm will likely be inefficient for analyzing these relationships, which is why other network based models such as TransE its derivatives, which specifically handles edge types differently and has been applied to biological data (85–88), are being explored.

In the networks used for this study, there is a complete lack of pharmacokinetics/ pharmacodynamics measures that can even better describe compound behavior. Despite the interactomic network still producing impressive results without this data, future work will explore dimensions such as drug dose, concentration over time and in different tissues, metabolites, etc. that ideally can eliminate unrealistic interactions from the analyses. The network is still capable of learning from a thermodynamic standpoint, but the introduction of separate nodes denoting the same drug at doses X and Y may eliminate fallacious candidates (such as ivermectin for COVID-19 (89, 90)).

Conclusions

This study is the first foray using the CANDO platform into large-scale integration of higher level biomedical processes and entities such as protein pathways, protein-protein interactions, adverse drug reactions, and protein-indication associations for the primary purpose of better understanding the totality of interactions, or the interactomic signatures, of drugs in relation to not only diseases they are known to treat, but the ADRs that they cause. The rich signal provided by the ADRs for describing drug behavior far outweighed the contributions of other relations in terms of drug-indication benchmarking, which will be further explored in future studies. However, the interactomic approach was still able to accurately predict drug-indication associations when drugs were considered by only their distance to a subset of the pathways in the network, indicating the contextual information of the protein interactions is still relevant for describing their behavior. Further, the networks constructed without any ADRs still outperformed the linear proteomic signature pipeline, albeit to a much lesser extent, yet provides further justification to pursue the interactomic approaches to continue to enhance how well we can describe drug behavior and works towards advancing the science of drug discovery.

ACKNOWLEDGMENTS. This work was supported in part by a NIH Director's Pioneer Award (DP1OD006779), a NIH Clinical and Translational Sciences Award (UL1TR001412), NIH T15 Award (T15LM012495), an NCATS ASPIRE Design Challenge Award, an NCATS ASPIRE Reduction-to-Practice Award, and startup funds from the Department of Biomedical Informatics at the University at Buffalo. The authors would like to acknowledge the support provided by the Center for Computational Research at the University at Buffalo. We would also like to thank all members of the Samudrala Computational Biology Group.

1. A Mullard, New drugs cost us [dollar] 2.6 billion to develop. *Nat. Rev. Drug Discov.* **13**, 877–877 (2014).
2. DW Thomas, et al., Clinical development success rates 2006–2015. *BIO Ind. Analysis* **1**, 16 (2016).

DRAFT

3. M Hay, DW Thomas, JL Craighead, C Economides, J Rosenthal, Clinical development success rates for investigational drugs. *Nat. biotechnology* **32**, 40–51 (2014).
4. CH Wong, KW Siah, AW Lo, Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).
5. J Eder, PL Herrling, Trends in modern drug discovery. *New Approaches to Drug Discov.* pp. 3–22 (2015).
6. M Deininger, Imatinib–an overview in *Hematology Meeting Reports (formerly Haematologica Reports).* Vol. 1, (2005).
7. GJ Gleich, KM Leiferman, A Pardanani, A Tefferi, JH Butterfield, Treatment of hypereosinophilic syndrome with imatinib mesilate. *The Lancet* **359**, 1577–1578 (2002).
8. H Joensuu, Treatment of inoperable gastrointestinal stromal tumor (gist) with imatinib (glivec, gleevec). *Medizinische Klinik (Munich, Ger. 1983)* **97**, 28–30 (2002).
9. HJ Droogendijk, et al., Imatinib mesylate in the treatment of systemic mastocytosis: a phase ii trial. *Cancer* **107**, 345–351 (2006).
10. SJ Lee, JY Wang, Exploiting the promiscuity of imatinib. *J. biology* **8**, 1–4 (2009).
11. GM Maggiora, The reductionist paradox: are the laws of chemistry and physics sufficient for the discovery of new drugs? *J. computer-aided molecular design* **25**, 699–708 (2011).
12. W Zheng, N Thorne, JC McKew, Phenotypic screens as a renewed approach for drug discovery. *Drug discovery today* **18**, 1067–1073 (2013).
13. H Akil, et al., Treatment resistant depression: a multi-scale, systems biology approach. *Neurosci. & Biobehav. Rev.* **84**, 272–288 (2018).
14. I Tavassoly, J Goldfarb, R Iyengar, Systems biology primer: the basic methods and approaches. *Essays biochemistry* **62**, 487–500 (2018).
15. EL Berg, Systems biology in drug discovery and development. *Drug discovery today* **19**, 113–125 (2014).
16. C Fotis, A Antoranz, D Hatziavramidis, T Sakellaropoulos, LG Alexopoulos, Network-based technologies for early drug discovery. *Drug discovery today* **23**, 626–635 (2018).
17. P Csermely, T Korcsmáros, HJ Kiss, G London, R Nussinov, Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. & therapeutics* **138**, 333–408 (2013).
18. I Wallach, M Dzamba, A Heifets, Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855* (2015).
19. K Li, Y Du, L Li, DQ Wei, Bioinformatics approaches for anti-cancer drug discovery. *Curr. drug targets* **21**, 3–17 (2020).
20. H Iwata, et al., Inferring protein domains associated with drug side effects based on drug-target interaction network. *BMC systems biology* **7**, S18 (2013).
21. T Liu, RB Altman, Relating essential proteins to drug side-effects using canonical component analysis: a structure-based approach. *J. chemical information modeling* **55**, 1483–1494 (2015).
22. Y Yamanishi, E Pauwels, M Kotera, Drug side-effect prediction based on the integration of chemical and biological spaces. *J. chemical information modeling* **52**, 3284–3292 (2012).
23. M Minie, et al., Cando and the infinite drug discovery frontier. *Drug discovery today* **19**, 1353–1363 (2014).
24. G Sethi, G Chopra, R Samudrala, Multiscale modelling of relationships between protein classes and drug behavior across all diseases using the cando platform. *Mini reviews medicinal chemistry* **15**, 705–717 (2015).
25. G Chopra, R Samudrala, Exploring polypharmacology in drug discovery and repurposing using the cando platform. *Curr. pharmaceutical design* **22**, 3109–3123 (2016).
26. Z Falls, W Mangione, J Schuler, R Samudrala, Exploration of interaction scoring criteria in the cando platform. *BMC Res. Notes* **12**, 318 (2019).
27. W Mangione, R Samudrala, Identifying protein features responsible for improved drug repurposing accuracies using the cando platform: Implications for drug design. *Molecules* **24**, 167 (2019).
28. J Schuler, R Samudrala, Fingerprinting cando: Increased accuracy with structure-and ligand-based shotgun drug repurposing. *ACS Omega* **4**, 17393–17403 (2019).
29. ML Hudson, R Samudrala, Multiscale virtual screening optimization for shotgun drug repurposing using the cando platform. *Molecules* **26**, 2581 (2021).
30. W Mangione, Z Falls, G Chopra, R Samudrala, cando.py: Open source software for predictive bioanalytics of large scale drug–protein–disease data. *J. chemical information modeling* **60**, 4131–4136 (2020).
31. G Chopra, S Kaushik, PL Elkin, R Samudrala, Combating ebola with repurposed therapeutics using the cando platform. *Molecules* **21**, 1537 (2016).
32. W Mangione, Z Falls, T Melendy, G Chopra, R Samudrala, Shotgun drug repurposing biotechnology to tackle epidemics and pandemics. *Drug Discov. Today* **25**, 1126–1128 (2020).
33. GO Consortium, The gene ontology (go) database and informatics resource. *Nucleic acids research* **32**, D258–D261 (2004).
34. GO Consortium, Gene ontology consortium: going forward. *Nucleic acids research* **43**, D1049–D1056 (2015).
35. GO Consortium, The gene ontology resource: 20 years and still going strong. *Nucleic acids research* **47**, D330–D338 (2019).
36. A Grover, J Leskovec, node2vec: Scalable feature learning for networks in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining.* pp. 855–864 (2016).
37. DS Wishart, et al., Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2017).
38. HM Berman, et al., *The protein data bank, 1999–.* (Springer), pp. 675–684 (2006).
39. SK Burley, et al., Rcsb protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic acids research* **47**, D464–D474 (2019).
40. Y Zhang, I-tasser server for protein 3d structure prediction. *BMC bioinformatics* **9**, 40 (2008).
41. D Xu, J Zhang, A Roy, Y Zhang, Automated protein structure modeling in casp9 by i-tasser pipeline combined with quark-based ab initio folding and fg-md-based structure refinement. *Proteins: Struct. Funct. Bioinforma.* **79**, 147–160 (2011).
42. J Yang, et al., The i-tasser suite: protein structure and function prediction. *Nat. methods* **12**, 7–8 (2015).
43. U Consortium, Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research* **47**, D506–D515 (2019).
44. JM Dana, et al., Sifts: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic acids research* **47**, D482–D489 (2019).
45. S Velankar, et al., Sifts: structure integration with function, taxonomy and sequences resource. *Nucleic acids research* **41**, D483–D489 (2012).
46. M Remmert, A Biegert, A Hauser, J Söding, Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nat. methods* **9**, 173–175 (2012).
47. AP Davis, et al., Comparative toxicogenomics database (ctd): update 2021. *Nucleic acids research* **49**, D1138–D1143 (2021).
48. J Piñero, et al., Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research* p. gkw943 (2016).
49. CE Lipscomb, Medical subject headings (mesh). *Bull. Med. Libr. Assoc.* **88**, 265 (2000).
50. M Kuhn, I Letunic, LJ Jensen, P Bork, The sider database of drugs and side effects. *Nucleic acids research* **44**, D1075–D1079 (2016).
51. NP Tatonetti, PY Patrick, R Daneshjou, RB Altman, Data-driven prediction of drug effects and interactions. *Sci. translational medicine* **4**, 125ra31–125ra31 (2012).
52. EG Brown, L Wood, S Wood, The medical dictionary for regulatory activities (meddra). *Drug safety* **20**, 109–117 (1999).
53. O Bodenreider, The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research* **32**, D267–D270 (2004).
54. K Donnelly, , et al., Snomed-ct: The advanced terminology and coding system for ehealth. *Stud. health technology informatics* **121**, 279 (2006).
55. B Jassal, et al., The reactome pathway knowledgebase. *Nucleic acids research* **48**, D498–D503 (2020).
56. D Szklarczyk, et al., String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research* **47**, D607–D613 (2019).
57. G Ayano, Dopamine: receptors, functions, synthesis, pathways, locations and mental disorders: review of literatures. *J Ment Disord Treat* **2**, 2 (2016).
58. Si Ueno, Genetic polymorphisms of serotonin and dopamine transporters in mental disorders. *J. Med. Investig.* **50**, 25–31 (2003).
59. HH Nickols, PJ Conn, Development of allosteric modulators of gpcrs for treatment of cns disorders. *Neurobiol. disease* **61**, 55–71 (2014).
60. H Komatsu, Novel therapeutic gpcrs for psychiatric disorders. *Int. journal molecular sciences* **16**, 14109–14121 (2015).
61. H Komatsu, M Fukuchi, Y Habata, Potential utility of biased gpcr signaling for treatment of psychiatric disorders. *Int. journal molecular sciences* **20**, 3207 (2019).
62. L Guglielmi, et al., Update on the implication of potassium channels in autism: K+ channelautism spectrum disorder. *Front. cellular neuroscience* **9**, 34 (2015).
63. W Noh, S Pak, G Choi, S Yang, S Yang, Transient potassium channels: Therapeutic targets for brain disorders. *Front. cellular neuroscience* **13**, 265 (2019).
64. JT Judy, P Zandi, A review of potassium channels in bipolar disorder. *Front. genetics* **4**, 105 (2013).
65. K Iwamoto, M Bundo, T Kato, Serotonin receptor 2c and mental disorders: genetic, expression, and rna editing studies. *RNA biology* **6**, 248–253 (2009).
66. A Haduch, WA Daniel, The engagement of brain cytochrome p450 in the metabolism of endogenous neuroactive substrates: a possible role in mental disorders. *Drug metabolism reviews* **50**, 415–429 (2018).
67. N Mahmood, A Arakelian, D Cheishvili, M Szyf, SA Rabbani, S-adenosylmethionine in combination with decitabine shows enhanced anti-cancer effects in repressing breast cancer growth and metastasis. *J. cellular molecular medicine* **24**, 10322–10337 (2020).
68. N Mahmood, et al., Methyl donor s-adenosylmethionine (sam) supplementation attenuates breast cancer growth, invasion, and metastasis in vivo; therapeutic and chemopreventive applications. *Oncotarget* **9**, 5169 (2018).
69. S Parashar, et al., S-adenosylmethionine blocks osteosarcoma cells proliferation and invasion in vitro and tumor metastasis in vivo: therapeutic and diagnostic clinical applications. *Cancer medicine* **4**, 732–744 (2015).
70. M Lavy, V Gauttier, N Poirier, S Barillé-Nion, C Blanquart, Specialized pro-resolving mediators mitigate cancer-related inflammation: role of tumor-associated macrophages and therapeutic opportunities. *Front. Immunol.* **12** (2021).
71. DR Vatnick, et al., Control of breast cancer through the resolution of inflammation. *The FASEB J.* **30**, 698–3 (2016).
72. CN Serhan, et al., Maresins: novel macrophage mediators with potent antiinflammatory and proresolving actions. *J. Exp. Medicine* **206**, 15–23 (2009).
73. DL Sudakin, Dietary aflatoxin exposure and chemoprevention of cancer: a clinical review. *J. Toxicol. Clin. Toxicol.* **41**, 195–204 (2003).

DRAFT

74. S Marchese, et al., Aflatoxin b1 and m1: Biological properties and their involvement in cancer development. *Toxins* **10**, 214 (2018).
75. PE Jackson, JD Groopman, Aflatoxin and liver cancer. *Best Pract. & Res. Clin. Gastroenterol.* **13**, 545–555 (1999).
76. CD Poff, M Balazy, Drugs that target lipoxygenases and leukotrienes as emerging therapies for asthma and cancer. *Curr. Drug Targets-Inflammation & Allergy* **3**, 19–33 (2004).
77. HE Claesson, On the biosynthesis and biological role of eoxins and 15-lipoxygenase-1 in airway inflammation and hodgkin lymphoma. *Prostaglandins & other lipid mediators* **89**, 120–125 (2009).
78. V Heyd, A Eynard, Effects of eicosatrienoic acid (20: 3 n-9, mead's acid) on some promalignant-related properties of three human cancer cell lines. *Prostaglandins & other lipid mediators* **71**, 177–188 (2003).
79. MK Apaya, PW Hsiao, YC Yang, LF Shyur, Deregulating the cyp2c19/epoxy-eicosatrienoic acid-associated fabp4/fabp5 signaling network as a therapeutic approach for metastatic triple-negative breast cancer. *Cancers* **12**, 199 (2020).
80. Y Ma, et al., Fatty acid oxidation: An emerging facet of metabolic transformation in cancer. *Cancer letters* **435**, 92–100 (2018).
81. Y Ma, et al., Long-chain fatty acyl-coa synthetase 1 promotes prostate cancer progression by elevation of lipogenesis and fatty acid beta-oxidation. *Oncogene* **40**, 1806–1820 (2021).
82. Y Liu, Fatty acid oxidation is a dominant bioenergetic pathway in prostate cancer. *Prostate cancer prostatic diseases* **9**, 230–234 (2006).
83. A Subramanian, et al., A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
84. J Lamb, et al., The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science* **313**, 1929–1935 (2006).
85. W Choi, H Lee, Identifying disease-gene associations using a convolutional neural network-based model by embedding a biological knowledge graph with entity descriptions. *Plos one* **16**, e0258626 (2021).
86. J Peng, G Lu, X Shang, A survey of network representation learning methods for link prediction in biological network. *Curr. pharmaceutical design* **26**, 3076–3084 (2020).
87. R Ye, X Li, Y Fang, H Zang, M Wang, A vectorized relational graph convolutional network for multi-relational network alignment. in *IJCAI*. pp. 4135–4141 (2019).
88. C Tu, Z Zhang, Z Liu, M Sun, Transnet: Translation-based network representation learning for social relation extraction. in *IJCAI*. pp. 2864–2870 (2017).
89. VD Schmith, J Zhou, LR Lohmer, The approved dose of ivermectin alone is not the ideal dose for the treatment of covid-19. *Clin. Pharmacol. & Ther.* **108**, 762–765 (2020).
90. G Momekov, D Momekova, Ivermectin as a potential covid-19 treatment from the pharmacokinetic point of view: antiviral levels are not likely attainable with known dosing regimens. *Biotechnol. & biotechnological equipment* **34**, 469–474 (2020).

DRAFT