Benchmarking causal reasoning algorithms for gene expression-based compound mechanism of action analysis

Layla Hosseini-Gerami
Centre for Molecular Informatics, Department of Chemistry, University of Cambridge
David A. Collier
Eli Lilly and Company
Emma Laing
Eli Lilly and Company
David Evans
Eli Lilly and Company
Howard Broughton
Eli Lilly and Company, Centre de Investigación, Alcobendas
Andreas Bender (ab454@cam.ac.uk)
Centre for Molecular Informatics, Department of Chemistry, University of Cambridge

Research Article

Keywords: Transcriptomics, Causal Reasoning, Mechanism of Action, Network Biology, Benchmarking, L1000

Posted Date: January 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1239049/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Bioinformatics on April 18th, 2023.
See the published version at https://doi.org/10.1186/s12859-023-05277-1.

Abstract

Background

Elucidating compound mechanism of action (MoA) is beneficial to drug discovery, but in practice often represents a significant challenge. Causal Reasoning approaches aim to address this situation by inferring dysregulated signalling proteins using transcriptomics data and biological networks; however, a comprehensive benchmarking of such approaches has not yet been reported. Here we benchmarked four causal reasoning algorithms (SigNet, CausalR, CausalR ScanR and CARNIVAL) with four networks (the smaller Omnipath network vs. 3 larger MetaBase™ networks), using LINCS L1000 and CMap microarray data, and assessed to what extent each factor dictated the successful recovery of direct targets and compound-associated signalling pathways in a benchmark dataset comprising 269 compounds. We additionally examined impact on performance in terms of the functions and roles of protein targets and their connectivity bias in the prior knowledge networks.

Results

According to ANOVA analysis, the combination of algorithm and network most significantly dictated the performance of causal reasoning algorithms, with the SigNet recovering the greatest number of *direct targets*. With respect to the recovery of *signalling pathways*, CARNIVAL with the Omnipath network was able to recover the most informative pathways containing compound targets, based on the Reactome pathway hierarchy. We found no significant difference in performance between L1000 data or microarray data, even when limited to just 978 'landmark' genes. Notably, all causal reasoning algorithms also outperformed pathway recovery based on input DEGs, despite these often being used for pathway enrichment. Causal reasoning methods performance was somewhat correlated with connectivity and biological role of the targets.

Conclusions

Overall, we conclude that causal reasoning performs well at recovering compound MoA upstream from gene expression changes by leveraging prior knowledge networks, and that the choice of network and algorithm has a profound impact on the performance of causal reasoning algorithms. Based on the analyses presented here this is true for both microarray-based gene expression data as well as those based on the L1000 platform.

Introduction

Following their administration in a biological system, compounds act to modulate the activity of signalling networks and pathways *via* (direct and indirect) interactions with protein targets [1]. The targets, signalling proteins and biological pathways modulated by the compound describe the

mechanism of action (MoA) of a compound. This MoA can also be defined by associated changes to cellular metabolism, transcriptional regulation and cell morphology[2], which may arise a direct consequence of compound treatment or *via* feedback loops, all of which contribute to a "systems-level" view, as opposed to the "one drug, one target" reductionist view of compound MoA[3]. Previous reviews have highlighted the importance of understanding compound MoA to guide drug discovery[4] – not only to validate observed phenotypic effects, but to understanding side effects[5], identify opportunities for personalised medicine[6], and to repurpose approved drugs for new indications[7]. The complex nature of a compound's MoA, and the fact that it can be defined by different layers of biology, means that for uncharacterised compounds the elucidation of its MoA is generally a significant bottleneck. To this end, chem- and bioinformatics approaches, based on different types of bioactivity and "-omics" data (such as genomics, proteomics, metabolomics, and – of direct relevance for the current work – transcriptomics), have become popular for generating testable hypotheses by harnessing experimental data with mathematical and statistical analyses and computational algorithms[8].

Recently, large -omics databases have become available in the public domain, such as the LINCS L1000 database[9] (scale-up of the Connectivity Map database[10] with a more high-throughput platform) which catalogues the transcriptional response of a variety of cell lines to treatment with ~30,000 different small molecules by measuring a reduced representation of the transcriptome (978 genes) and inferring the expression levels of the remaining genes from this so-called "landmark" set. Although these large data sets provide a rich *starting point* for the understanding of drug mechanisms, a key question is *how* such data should be exploited and interpreted in a meaningful way to shortlist targets or pathways for experimental validation. Transcriptomics experiments allow for the identification of genes which are significantly differentially expressed upon compound treatment compared to a control (DEGs, or differentially expressed genes), but can produce long lists of genes which are difficult to interpret in terms of mechanistic hypotheses without further analysis[11]. Common strategies for interpretation of these data include machine learning using gene expression measurements as features to predict MoA labels[12], although such approaches are limited due to the challenges in labelling compounds with "gold-standard" MoA annotations[13, 14] and are often given quite shallow and broad labels, e.g. "dopamine receptor inhibitor", which do not inform about receptor subtype information, as well as any downstream pathways being affected which would influence the phenotypic effects of the compound. Alternatively, DEGs can be analysed with pathway enrichment methods which calculate the statistical significance of the association of their protein products with annotated biological pathways and processes[15]. Although this approach provides a simple way to reduce large gene sets down to a smaller set of biologically interpretable pathways, it relies on the only partially true association of gene expression with protein activity and abundance[11]. In fact, protein activity is dictated not only by transcription, but translation and post-translational modifications such as phosphorylation – differential gene expression has indeed been found to reflect the activity of upstream transcription factors rather than the activity of a pathway of interest[16].

Methods known collectively as "causal reasoning" have been developed with the aim of identifying causal molecules (of an observed response) by treating differential gene expression as a *consequence* of

differential protein activity, rather than equating gene expression with signalling protein activity. Such methods maximise the biological information gained with transcriptomics data by incorporating prior knowledge networks (PKNs) of signed and directed (i.e., *X* inhibit*s Y*) protein-protein interactions (PPIs) to trace upstream of mRNA regulation to the targets and signalling proteins modulated by a compound (Figure 1).

Because transcription factors (TFs) are modulated due to an external stimulus such as compound binding to its target receptor(s), subsequently leading to cellular signal transduction and modulation of TF activity[17] (Figure 1A), gene expression data can be linked back, using causal reasoning, to the upstream proteins on a PKN that are most likely to be responsible for the observed (measured) transcriptional changes using causal reasoning. These causal proteins could represent inferred signalling proteins or compound targets, and can thus be viewed in terms of compound mechanism of action, and represent experimentally testable hypotheses (Figure 1B).

Case studies of mechanism of action elucidation using causal reasoning include the elucidation of key processes involved in a DGAT1 inhibitor for obesity[18], and an AKT inhibitor for cancer[19], with the former study remarking that the inferred signalling proteins represented less high-level and more detailed processes in contrast to the findings derived from traditional pathway enrichment methods. This illustrates the utility of causal reasoning methodologies to derive mechanistic hypotheses from transcriptomics data, overcoming the limitations of other related methods discussed above. However, to date, no large-scale benchmarking study to systematically investigate different parameters affecting the ability of different causal reasoning algorithms to characterise compound mechanism of action in terms of different levels of biology (i.e., direct targets and modulated pathways) has been reported, and such a study would hence be of interest to researchers looking to use such methods.

One previous benchmarking study[20] compared several computational network algorithms, finding that the causal reasoning algorithm SigNet[21] (also considered in our study) performed relatively well at recovering target proteins from compound-perturbed gene expression data (from the Connectivity Map dataset[10]), ranking in the top 5 of 17 algorithms in ~35% of cases, in terms of fraction of direct targets recovered. The authors did however not consider the impact of prior knowledge network, and neither did they investigate the applicability domain with respect to the nature of protein classes and network connectivity bias and (see below) only considered direct targets. In terms of network benchmarking, a study which aimed to compare the ability to infer disease-related genes from a subset of disease genes found that performance scaled with network size (Pearson's r = 0.88 for number of interactions in the network vs. its overall average rank in recovery of disease genes), indicating that the benefits of adding more information outweighed potential detrimental effects arising from noise (e.g. false positive interactions, or more distantly related information)[22].

The aforementioned algorithm benchmarking study[20] only focused on the ability of the algorithms to retrieve *direct* targets, rather than considering the signalling proteins and pathways. Although protein targets are one way to define compound mechanism of action[23], the inference of particular targets from

gene expression data is a difficult task due to both conceptual and practical limitations. On the conceptual side, not all target modulations lead to downstream effects in gene expression[24], and gene expression data is downstream of the modulation of often multiple targets[25]. On the practical side, there are high levels of noise in both gene expression data[26] and biological networks[27], and bioactivity datasets are sparse (92% sparsity according to a previous study of PubChem and ChEMBL data[28]). In addition, causally inferred signalling proteins can be associated with biological processes and pathways[29] and related to the compound's mechanism of action to obtain an understanding of compound mechanism of action on a deeper level compared to solely focusing on direct target retrieval. Hence, in our study we sought to measure the success of causal reasoning algorithms both in terms of recovered targets *and* pathways.

Current Study

This study aimed to benchmark casual reasoning algorithms for their ability to recover compound mechanism of action (on both the target- *and* pathway-level) from gene expression data, in particular with respect to the following factors which we anticipated would present the greatest influence on the results:

- Algorithm scoring methodology (CausalR, ScanR, SigNet, CARNIVAL)
- Source of input gene expression data (CMap or LINCS, MCF7 or PC3 cell line)
- Input gene set (landmark genes, landmark and best inferred, or all)
- Prior knowledge network (smaller Omnipath network vs. 3 larger MetaBase™ networks)
- Properties and biological functions of protein nodes (connectivity on network, protein class)

In terms of causal reasoning algorithms, we investigated the CausalR[30] (both node ranking and ScanR subnetwork outputs), CARNIVAL[29] (subnetwork output) and SigNet[21] (node ranking) methodologies. CausalR measures concordance and significance of a protein's known interactions with the observed transcriptional changes, and allows one to "scan" multiple path lengths to find consensus inferred regulator proteins (ScanR) and link them to the downstream measurements *via* concordant interactions, CARNIVAL aims to optimise a network of inferred signalling proteins with Integer Linear Programming and weighting of network nodes *a priori* using the gene expression pathway "footprint", and SigNet takes the consensus of multiple scoring algorithms which deal with network topology in different ways and additionally takes gene-level measurements into account.

In terms of input gene expression data we made use of the publicly available CMap[10] and LINCS L1000[9] gene expression databases for the compound signatures used in the study. Compounds were selected based on their presence in both databases, measured in the MCF7 and PC3 cell lines at 10μM and after 6 hours of treatment, and with known target annotations in the public domain.

By comparing L1000 with its predecessor, CMap, we aimed to disentangle any effects arising from the *quantity* of input genes (i.e., using 978 instead of 10,000+) from the effects arising from the *quality* of

poorly-predicted non-landmark genes, as well as any differences in technologies between the two platforms. Hence, for each compound measured under particular conditions (cell line, time point, dose), we obtained three signatures forming three different input gene sets – consisting of landmark genes only (as defined by the L1000 study), “best-inferred” genes (again defined by the L1000 study as genes which show the highest accuracy of inference), and all genes.

We also evaluated whether different algorithms handled increasing network size and density in a different manner, owing to the differences in how they prioritise causal proteins, and to understand whether performance scales with network size. Hence, we used the publicly available Omnipath network (3,610 nodes and 9,306 edges), as well as commercial MetaBase™ networks in our study at 3 confidence levels (up to 11,004 nodes and 87,556 edges) (Supplementary Table 1).

To quantify the performance of the algorithms, we assessed the ability of each algorithm, in combination with different networks and parameter settings, to recover known targets directly, as well as compound-associated pathways (those annotated with the known targets) using a two-step enrichment approached described in the CARNIVAL study[29]. We obtained known compound targets from ChEMBL[31], RepurposingHub[32] and Connectivity Map[10] to allow us to evaluate the output of the algorithms. It is also important to note that annotated targets may not directly relate to compound efficacy, for example in the case of off-targets, and *in vitro* activities in particular may not be relevant for *in vivo* mechanism of action (for example due to PK; and indeed, cancer cell lines may not necessarily reflect *in vivo* gene expression response), however for the purpose of this study we use target annotations as the best proxy for mechanism of action which is currently available in the public domain for a large-scale benchmark study. Following the computation of both evaluation metrics, we used a full-factorial fixed effects ANOVA model to understand the contribution of each parameter (input data, network, algorithm) to each evaluation metric, evaluated the applicability domain of the methods by investigating any potential association with the successful recovery of a particular target class and its connectivity in the prior knowledge networks, or its biological role. We additionally benchmarked the results against baseline results derived from input gene expression data, where differentially expressed genes were used in place of causal nodes.

Hence, the overall aim of this study was to systematically investigate the ability of causal reasoning algorithms to infer compound mechanism of action on both the target- and pathway-levels, with respect to both the key factors which influence their quantitative performance as well as for which targets/signalling proteins the methods are likely to be more or less successful.

Materials And Methods

The overall workflow for this study can be found in Figure 2.

Transcriptomics data extraction

(Figure 2A) L1000 data (Level 5) were extracted (access date September 2019) using the CMapPy [33] package from .gctx files found at GEO accessions *GSE92742* (Phase I) and *GSE70138* (Phase II) [9], with a Python script available on GitHub (https://github.com/BenderGroup/LINCS-Extraction). All compound data derived at a time point of 6h after compound treatment and using a compound concentration of 10μM were extracted for this study, measured in the two cell lines common to CMap and LINCS (MCF7 and PC3). While this (relatively high) concentration is unlikely to represent physiological compound concentrations after drug administration, here only a link on the *in vitro* level between gene expression signal and compound mode of action was aimed to be established, plus the majority of data was only available at this concentration from the practical angle. In cases where compounds possessed biological replicates at level 5 (which represents differential gene expression signatures), replicate signatures were weighted based on their co-correlation and combined linearly with their weights to produce a consensus signature (as described for technical replicates in the L1000 protocol[9], Python code available from the aforementioned GitHub link). CMap data was extracted using the protocol and functions described in longevityTools R package eDRUG vignette[34], which included the extraction of CEL files from the CMap website (https://portals.broadinstitute.org/cmap/), normalisation with MAS5[35] and differential analysis with *limma*[36]. To investigate the effect of including inferred transcriptomics data on mechanism of action recovery, both the LINCS and CMap data were separated into three subsets – landmark genes only (based on the definition of landmark genes from the L1000 study found in the “gene_info.txt” metadata), landmark and “best inferred genes” (again based on the L1000 metadata), and landmark and all inferred genes. The rationale for using subsets of CMap genes as well as L1000 was to separate the effects of quantity (landmark genes, vs all genes) and technology of derivation of gene expression data (microarray data vs L1000 platform data) on the results.

Compound target annotation

LINCS “pert_info.txt” metadata files were used to map compound preferred name (pert_iname) to InChI key. InChI keys for compounds were queried *via* the ChEMBL 25[31] mySQL database, with any protein bioactivity measurement corresponding to an activity of 10μM (pChEMBL of >=5) or better being annotated to the compound as a target. Target data were also extracted from the clue.io Repurposing Hub (2019) and Connectivity Map [32] using InChI keys. All annotations were merged together by InChI key (2016 compounds in total), with compounds with no annotations dropped. This formed a set of 269 compounds with bioactivity data and transcriptomic data measured in both LINCS and CMap in the MCF7 and PC3 cell lines (Additional File 1, Figure 2B). Disease Area annotations were also obtained from clue.io for the 269 compounds, where available. Targets were mapped to their corresponding protein classes using the ChEMBL webresource client (v. 0.10.0). A summary of the proportion of Disease Area annotations and target protein classes represented in the set of 269 compounds can be seen in Supplementary Figure 1.

Extraction of prior knowledge networks

For this study, we considered networks from Omnipath and MetaBase™ (Figure 2C). The Omnipath[37] protein-protein interaction (PPI) network (signed and directed edges only) was extracted from the

CARNIVAL[29] supplementary data with nodes as UniProt IDs. The IDs were converted to HGNC symbols using the *org.Hs.eg.db* package (v. 3.8.2)[38]. For SigNet, additional transcriptional regulatory interactions were extracted from OmnipathDB using the 'tf_target' and 'dorothea' filters, and concatenated with the PPI network (accessed November 2019).

The entire MetaBase™ network was first extracted using the get.globalnetwork(species="human") function of metabaseR (v. 4.2.3, November 2019, Thomspon Reuters^{TM}). Edges that were not direct molecular interactions were removed. Network object IDs were then converted to their corresponding Entrez IDs using the CBDD (November 2019, Thompson Reuters^{TM}) convertNetworkObjects2Entrez(networkobject,species="human") function. Nodes which did not convert to Entrez IDs were removed. If one network object corresponded to multiple Entrez IDs – i.e. a protein complex or a family – it was assumed that all members of the protein complex take part in the interaction. Only signed and directed edges were kept. To retain only protein-protein interactions from the global network, interactions annotated with "co-regulation of transcription", "Transcription regulation", "Influence on expression", "miRNA binding", "Unspecified", or "Pharmacological effect" were removed. For SigNet, to retain protein-protein and transcriptional regulatory interactions, only the interactions with mechanism "miRNA binding", "Unspecified" and "Pharmacological effect" were removed. Each network was then split into three confidence levels based on the "trust" attribute of each interaction (low, medium, high) – high confidence interactions only, high and medium interactions, and all interactions including low, medium and high confidence.

Transcriptomic signature preparation and processing (for CARNIVAL and CausalR)

For CARNIVAL and CausalR, signatures were processed further to yield transcription factor (TF) activities and pathway scores (these steps were not necessary for SigNet as the algorithm uses transcriptional regulation interactions in the prior knowledge network to link gene expression data to TF activities as the first step). For the CMap data, genes were input with their log2 fold-change values into DoRothEA, which uses a consensus TF-gene regulon to compute enrichment scores for transcription factors[39]. For the LINCS data, gene Z-scores were input into DoRothEA. As additional input to CARNIVAL, the same process was repeated with the PROGENy pipeline[40] to obtain pathway scores derived from a common core of pathway responsive genes, generated from a large compendium of perturbation experiments.

Causal Reasoning

For this benchmarking study we considered three different R packages, SigNet[21] (implemented in CBDD) CausalR[30] and CARNIVAL[29] (Figure 2D). Their inputs and outputs as well as a brief overview of their implemented algorithm(s) are summarised in Table 1.

CausalR (v. 1.16.0) was run with DoRothEA TF activities with the ReadExperimentalData() function. PPI networks were converted to computational causal graphs using the CreateCCG() function with default parameters. To carry out the causal reasoning, the ScanR() function was modified slightly and instead of

only considering the top N ranked regulators, searches for regulators which have a significant (p =< 0.05) score across all path lengths (up to a maximum of 5). If no regulators were found, path lengths 1 – n-1 were searched, and so on (R code available from https://github.com/laylagerami/CausalReasoningBenchmark/). Subnetworks were reconstructed for all consensus regulators including all correctly scored (concordant with the interactions in the PPI network) interactions from regulator to input TF, and aggregated to form an overall reconstructed compound-induced signalling network (herein referenced as 'Subnetwork' output). In addition, the ranked list of proteins at path length 5 was considered as an output to compare with the SigNet ranked list output (RankTheHypotheses(ccg,experimentalData,delta=5)) (herein referenced as 'Results Table' output).

The CARNIVAL (v. 0.9.3) *InvCarnival* algorithm (where known targets are not supplied as input) was run on the TF activities from DoRothEA and pathway scores from PROGENy using the runCARNIVAL() function with a time limit of 900 seconds and all other parameters kept as their default. The OmniPath and MetaBase™ PPIs were used as prior knowledge networks. The final pooled consensus subnetwork was taken as output.

The SigNet algorithm was applied to all datasets with the Log2FC (CMap) or Z-score (LINCS) used as input using the PPI networks with the additional transcriptional regulation interactions. The algorithmwas run using the SigNet() function in the *CBDD* package. All parameters were kept as default (path length of 5), and the ranked table was used as output. Directionality was not considered in thitudy, and ranked tables were concatenated so as to only consist of the highest ranked sign of each protein.

Table 1
The three algorithms considered in this benchmarking study

Algorithm	Input	Output	Summary
SigNet [21]	Gene expression PPI and transcriptional regulation network	Ranked list of network nodes	Gene expression mapped to TF activity using a TF-gene interactome, propagated through the protein-protein interaction network using a combination of scoring functions which compute the concordance of signalling proteins with the gene expression data
CausalR [30]	Transcription factor activities PPI network	Ranked list of network nodes Compound-specific subnetwork	Each network node scored based on the concordance of the protein's interactions with the input data. Scores are assessed for significance to account for uneven node connectivity ('hub nodes') which can often bias downstream results [41]. Frequently high scoring proteins across several path lengths can be combined with their concordant interactions ("ScanR") to generate a subnetwork of dysregulated proteins.
CARNIVAL [29]	Transcription factor activities Pathway activities (based on gene expression "footprint") PPI network	Compound-specific subnetwork	Integer linear programming along with pathways computed from the gene expression 'footprint' [40] to optimise a subnetwork of dysregulated signalling relating to the input data, which for CARNIVAL is required to be transcription factor activities inferred from the gene-level statistics from transcriptomics data [39].

Computation of evaluation metrics

Because the algorithms produced two types of output, subnetworks and ranked lists, not all evaluation metrics were able to be applied to every output type as summarized in Table 2. As SigNet does not compute significance values for ranked nodes, we computed the mean number of output nodes for the other algorithms; this was found to be 198, hence we used this as the cut-off for the top ranked nodes taken from the SigNet output.

Table 2
Algorithm outputs and the evaluation metrics applied in every case

Algorithm	Output	Target Recovery	Target Enrichment Factor	Pathway Recovery
CARNIVAL	Default output subnetwork	Y	N	Y
CausalR	ScanR subnetwork	Y	N	Y
CausalR	RankTheHypotheses table	Y Nodes with p<=0.05	Y	Y Nodes with p<=0.05
SigNet	Default output table (overall score)	Y Top 198 nodes	Y	Y Top 198 nodes

1. Target Recovery

To evaluate if direct targets were recovered, output from the algorithms was intersected with the known targets of each compound extracted from ChEMBL and RepurposingHub (Figure 2E). To compute the significance of the number of targets recovered based on the number of compound targets, the number of potential nodes recoverable from the prior knowledge network, and the cardinality of the output, the R function fishers.test(alternative=”greater”) was run. The number of recovered targets per compound were subsequently modelled in the ANOVA analysis. As a direct comparison to the input gene expression data, the transcriptomic data were subject to the same analysis where CMap genes with |Log2FC| >= 1.5 and LINCS genes |ZScore| >= 2 were compared to the targets. In this way we could understand if direct compound targets are often differentially regulated on the transcriptional level (which has been the results of previous analyses[42]). If the target was not present in the prior knowledge network (causal reasoning) or in the set of measured genes (LINCS/CMap), it was discounted from the significance calculation. Compound outputs where no targets could be recovered (i.e., no targets were present in the network) were discounted from the ANOVA statistical analysis. Supplementary Table 1 summarises target coverage for each network.

2. Target Enrichment Factor

To allow for further interpretation into the effect of the input network chosen, the Enrichment Factor was computed on the SigNet and CausalR ranked tables, where the true compound targets were designated as “true positives” and all other network nodes as “false positives” (although this might not be strictly true due to incomplete data being available), and each protein was ranked based on its score, using the enrichvs (v0.0.5) R package [43][44]. The Enrichment Factor was subsequently modelled in the ANOVA analysis.

3. Pathway Two-Step Enrichment

To quantify recovery of 'relevant pathways', we used the same principles as in Liu *et al*'s CARNIVAL evaluation metric which examines whether pathways which contain the target of interest are recovered[29], but instead combining *all* target-associated pathways into one set of compound-attributed pathways to take into account polypharmacology (Figure 2E). For each compound, each target was annotated with its participating pathways using the ReactomePA[45] R package. The set of unique target-annotated pathways was denoted as the compound-associated pathway set. Secondly, enrichment of output nodes in the Reactome pathway set was performed using the ReactomePA package[45] with all prior knowledge network nodes as the background/universe, and enrichment p-values adjusted with the Benjamini-Hochberg procedure[46]. This led to two pathway lists: Firstly, compound-associated pathways obtained from target annotations (ground truth), and, secondly, significantly enriched (adjusted p-value <= 0.05) pathways obtained from causal reasoning output nodes. To compute the over-representation of enriched pathways in the set of ground truth pathways, we computed a second enrichment p-value using fisher.test(alternative="greater") with all Reactome pathways set as the background/universe. We hence interpret the *second* enrichment p-value as the extent to which target-associated pathways were captured in the causal reasoning output, and this metric was subsequently modelled in the ANOVA analysis.

Furthermore, as Reactome pathways are laid out hierarchically under high-level categories such as "Cell Cycle", "Metabolism", "Neuronal System", we used the position of recovered pathways in the hierarchy, as well as the number of protein annotations, as a proxy for pathway specificity, where higher-level and larger pathway sets were assumed to be less specific and hence less informative for understanding compound mechanism of action. To this end, we downloaded the entire Reactome knowledgebase as a Neo4J object[47]. Following the initialisation of a Neo4J Reactome database, we retrieved the number of superpathways for each Reactome pathway (pathways above the pathway in question in the hierarchy) using the neo4r R package[48], with the command 'MATCH (p:Pathway{stId:[PATHWAY-ID]})<-[:hasEvent*]-(sp:Pathway) RETURN p.stId AS Pathway, sp.stId AS SuperPathway, sp.displayName as DisplayName'.

We again repeated the two-step enrichment analysis with pathways enriched from CMap genes |Log2FC| >= 1.5 and LINCs genes |ZScore| >= 2 using the same methodology. Compounds where no targets had annotated pathways were discounted from the ANOVA statistical analysis. If no pathways were significantly enriched, or if there were no genes with which to perform the enrichment, the second enrichment p-value was set to 1.

Statistical analysis

Following the computation of the aforementioned target and pathway recovery metrics, full-factorial fixed effects models were formulated so as to understand the contribution of each parameter (network, algorithm, platform, gene set and cell line) to the metric's score (Figure 2F). The ANOVA model for each metric was formulated as:

$$Metric \sim (Network + Algorithm + Platform + Gene\ Set + Cell\ Line)^2$$

And executed with the aov() R function across all results. Post-hoc least square means testing was carried out on main effects which were not included in interaction effects, and for interaction effects, in a pairwise fashion with Tukey-adjusted p-values. Least square means were weighted based on the frequencies in the original data of the factor combinations that were averaged over. An example formula for the Network:Algorithm interaction effect is:

Formula = pairwise ~ Network:Algorithm

This was achieved with the lsmeans(model, formula, adjust= “tukey”, weights = “proportional”) function from the emmeans R package[49].

Results And Discussion

Target and pathway recovery depends on network and algorithm

To understand which factors are most important when employing causal reasoning algorithms to elucidate compound mechanism of action, we performed full-factorial fixed effects ANOVA (see *Materials and Methods* for full details) modelling of each evaluation metric (illustrated in Figure 2) as a dependent variable and the parameters as regressors (network, algorithm, platform, gene set, and cell line). Table 3 displays the significant ($p < 0.05$) factors and interaction effects for each evaluation metric. For both evaluation metrics the Network and Algorithm factors as well as the Network:Algorithm interaction effect were statistically significant (Table 3, Supplementary Tables 2 & 3 for the full ANOVA results). Additionally, for the recovery of compound-associated pathways, the Algorithm:Gene set interaction effect was found to be statistically significant ($p = 0.00129$), however through examining the resulting interaction diagram we found that the effect was not practically relevant (only a significant but small performance increase using landmark genes vs. all genes with CARNIVAL, Supplementary Figure 2), hence we did not investigate this in more detail.

Table 3
Table of significant (p<0.05) factors and interaction effects for each evaluation metric, from ANOVA analysis

Metric	Factor(s)	p-value
Direct target recovery	Network	7.1×10^{-6}
	Algorithm	$< 2 \times 10^{-16}$
	Network:Algorithm	$< 2 \times 10^{-16}$
Compound-associated pathway recovery	Network	$< 2 \times 10^{-16}$
	Algorithm	$< 2 \times 10^{-16}$
	Network:Algorithm	$< 2 \times 10^{-16}$
	Algorithm:Gene set	0.00129

To next understand exactly how Network:Algorithm interact, we employed post-hoc least square means tests – computing the mean metric scores for every combination of network and algorithm, and algorithm and gene set, adjusted for the means of all other factors in the models. In this way we were able to discover how, and to what extent, the networks affected the ability of each algorithm to recover direct targets or compound-associated pathways. The resulting interaction diagram for the target recovery metric can be found in Figure 3, and the full results table from the least square means analysis can be seen in Supplementary Table 4. From the interaction plot in Figure 3 it can be seen that the use of the smaller, less dense Omnipath network led to a higher mean target recovery for both the CARNIVAL (0.37 targets per compound, 6.14% significant) and SigNet (1.38 targets per compound, 23.36% significant) algorithms compared to using the larger MetaBase™ networks (for each of “High”, “Medium” and “All” confidence levels), in terms of both the overlap and the percentage of cases where this overlap was significant (p<=0.05, Fisher’s Exact Test, when considering the number of targets, size of network and number of nodes recovered). While SigNet retained the highest percentage of significant (>20% for all networks), the CARNIVAL-recovered direct targets were more statistically significant with Omnipath compared to the MetaBase™ networks (6.14% vs. 3.21 and 3.4%). Conversely, the two CausalR algorithms recovered more compound targets directly with the larger, denser MetaBase™ networks, achieving a mean target overlap of just 0.16 (CausalR Results Table) and 0.31 (CausalR Subnetwork) targets per compound with Omnipath, lower than the gene expression DEG-target overlap baseline (0.38), with the percentage significant cases also decreasing. We did not observe any significant difference in performance between the different confidence MetaBase™ networks. The improved performance of CARNIVAL with Omnipath compared to the larger MetaBase™ networks for recovering direct targets is likely related to CARNIVAL being optimised using the same Omnipath network used in this study[29]. The results also indicate that SigNet tends to highly rank true positives using the smaller Omnipath network, whereas CausalR works more effectively when there are more interactions to “reason” over.

To investigate this finding further, we also computed the Enrichment Factor of targets for each ranked list output (SigNet and CausalR results table) within the top 5%, and found that indeed, the SigNet scoring function ranks true positives more highly with Omnipath compared to MetaBase™ networks (Enrichment Factor of 6 vs. ~3), and vice versa for CausalR (Supplementary Figure 3). This is in contrast to previous literature which suggests that the potential negative effects of noise are outweighed by a comprehensive inclusion of interactions when using biological networks as prior knowledge for node prioritisation algorithms [22, 50]. In fact, we can conclude from our analysis that - when using causal reasoning algorithms to prioritise compound targets – this behaviour is wholly dependent on the algorithm being used; the CausalR algorithms do benefit from a large prior knowledge network, whereas SigNet and CARNIVAL recover compound targets more effectively with the smaller Omnipath network which contains less potential noise or false positive interactions.

Overall, we found that SigNet with Omnipath showed the best performance with regard to recovering compound MoA in terms of direct targets (Figure 3). This performance was far higher than the gene expression baseline, indicating that differentially expressed genes only infrequently correspond to modulated targets. This general finding is in agreement with previous studies; Jaegar et al found that 15% of the time, and Iskar et al 8% of the time, did drug target expression level change upon drug treatment[21],[42]. The best individual overlap of DEGs with targets was 1.1 targets per compound, only 9.3% of overlaps being statistically significant ($p <= 0.05$), with all genes measured in the PC3 cell line using CMap data (Supplementary Figure 5). Hence, causal reasoning approaches such as SigNet provide a way to better relate downstream transcriptional changes to upstream target engagement than solely looking at DEGs themselves.

When considering the recovery of compound-target associated pathways (Figure 4A, full least square means analysis results in Supplementary Table 5), pathways derived from causal nodes from SigNet with the Ominpath network had the most significant least square mean enrichment (-log10(p-value)) of compound-associated pathways, with an enrichment value of 16.8, with the worst overall performance seen for the CausalR results table (and enrichment values between 1.02 and 2.08). The CausalR network output achieved a better performance over the results table (enrichment values between 7.91 and 8.83), meaning that CausalR does not highly rank direct compound protein targets, but it is able to find intermediate connecting nodes which represent proteins participating in target-annotated signalling pathways, and hence are potentially related to compound mechanism of action. Additionally, the results do not indicate that any one network tested has a significantly higher performance for the CausalR subnetwork results, according to the pairwise mean comparisons in Figure 4A. Like SigNet, CARNIVAL showed a significantly higher performance (enrichment value of 6.48) with the Omnipath network compared to the MetaBase™ networks (enrichment values of 2.19, 2.85 and 3.07 for Medium, High, and All, respectively) Notably, all least square mean enrichment results improved on the average gene expression baseline (enrichment value of 0.66), meaning that the Causal Reasoning algorithms were better able to recover relevant signalling proteins compared to using differentially expressed genes alone as a proxy for modulated signalling mediators. The best individual baseline (DEG) performance using gene expression data was using the CMap data, all genes, in the PC3 cell line, with a mean enrichment

value of 2.4 (Supplementary Figure 6) – in agreement with Liu et al's findings that CARNIVAL outperformed DEG enrichment for recovering relevant pathways[29].

As well as considering the enrichment p-value as a performance measure, we next aimed to understand how informative, and hence practically useful for understanding compound mechanism, the recovered pathways were. To quantify this, we assumed that pathways which are lower down in the Reactome pathway hierarchy (e.g., the high-level *Cell Cycle* pathway vs. the lower-level *Stabilization of p53* pathway), and that contain fewer genes/proteins, are more informative than higher-level, larger pathways. Figure 4B and 4C show the mean results for recovered pathway size (number of attributed proteins/genes), and hierarchy (number of superpathways), respectively. It was found that, despite the good performance in terms of statistical significance, SigNet with Omnipath pathways were both larger (300 genes on average) and higher up in the Reactome hierarchy (3.6 superpathways on average) compared to the MetaBase™ networks (gene set size below 160, number of superpathways greater than 4). CARNIVAL also recovered pathways that were larger (225 genes on average), but lower down in the Reactome hierarchy (4.9 superpathways on average), with Omnipath compared to the MetaBase™ networks (gene set size below 160, number of superpathways about 4.5 on average). Pathways recovered by the CausalR results table were the least informative, with the lowest number of superpathways on average, while the CausalR subnetwork recovered pathways had better results in comparison (smaller gene set sizes and greater number of superpathways), showing that the subnetwork methodology is superior to the results table when using CausalR to retrieve compound-associated pathways. Other than with respect to average pathway size for SigNet with Omnipath, all combinations of network and algorithm recovered more informative pathways compared to average baseline gene expression results (gene set size = 280, number of superpathways = 1.9), indicating that pathways recovered from DEGs generally capture higher-level processes compared to causally inferred proteins. This was also found in an application of causal reasoning to understand the mechanisms of a DGAT1 inhibitor, where the authors found that enrichment with gene expression data generally pointed to higher-level processes[18]. We believe that this is due to transcriptional changes capturing the *effect* of protein signalling, while causal nodes represent the signalling proteins themselves.

In general, the causal reasoning methodologies outperformed the DEG baseline results when deriving compound mechanism of action from gene expression data. We also found that, based on our experimental setup, the algorithms are generally robust with respect to the choice of transcriptomics platform, gene set and cell line – with these not having any significant effect on the performance metrics based on our full factorial fixed-effects models, except for the Algorithm: Gene set finding which only had a marginal influence with the CARNIVAL algorithm. Our study hence indicates that the LINCS L1000 data is therefore suitable for use with causal reasoning algorithms, at least based on the other factors used in this work, and that the 978 landmark genes are truly informative enough to gain insight into compound mechanism of action, as hypothesised in the original publication[9]. We do note however that the cell lines used in this study (PC3 and MCF7, chosen as they were also present in the original CMap), are transcriptionally similar on the baseline level (Supplementary Figure 7), and the use of other cell lines (e.g., non-cancer, or indeed *in vivo* models) is likely to alter performance.

Target recovery with Causal Reasoning is dependent on network bias and biological function

We next investigated performance of the algorithms as a function of the connectivity of a protein target in the network, due to the known connectivity bias present in biological networks[51], as well as its biological function.

For each combination of network and algorithm we calculated how many times each target was recovered. We then normalised this value to account for annotation prevalence by dividing it by the number of times the target was annotated in the compound set. Finally, we calculated the Spearman rank correlation of the normalised target recovery with their degree in the corresponding PPI network (Figure 5). The lowest correlation can be seen for the CausalR results table (mean of -0.04) which is likely due to the fact that each protein is given a significance value to explicitly correct for the known connectivity bias, and we take as output only the nodes with p <= 0.05. The correlation between target recovery and network connectivity was highest using the CausalR subnetworks (mean of 0.72) – an explanation for this is that they connect key drivers to input TFs through correctly explained interactions, and will therefore go through “hub” nodes more often. Despite this large difference in correlation, the two CausalR outputs performed roughly similarly in terms of direct target recovery (Figure 3), which indicates that the CausalR ranked table is better able to prioritise less-connected (and hence less-studied) targets compared to the subnetwork output. SigNet (mean of 0.21) and CARNIVAL (mean of 0.23) showed roughly similar correlation patterns, with the correlation between target recovery and network connectivity increasing (to 0.38 and 0.56, respectively) with use of the Omnipath network, corresponding to their increased performance with this network (Figure 3). Additionally, we found that targets recovered with the Omnipath network showed a higher correlation with network connectivity (mean of 0.42) compared to the MetaBase™ networks (means of 0.25, 0.24, 0.24), this is potentially due to the small size of the network making hub effects more prominent. Overall, target recovery performance with causal reasoning is generally associated with network connectivity – excluding the targets recovered with the CausalR results table. We further examined the degree distributions of targets vs non-targets on each network, finding that drug targets were more often found to have a higher connectivity on each network compared to non-targets (Supplementary Figure 8, and generally seen in previous studies[52]). Therefore, a high correlation between target recovery and network connectivity is not necessarily detrimental - however, this bias would affect the recovery of less-studied proteins which must be kept in mind depending on the disease area being studied.

We next analysed the best-performing algorithm (namely, SigNet with the Omnipath network) in more detail with respect to its ability to recover targets across different protein classes (annotations retrieved from ChEMBL), comparing the findings with SigNet with the full MetaBase™ network, the results of which are shown in Figure 6. We computed protein class recovery in the same way as target recovery, calculating how many times a target in protein class was recovered and normalising this value by the annotation prevalence of the protein class in the compound set, converting this value to an overall

percentage. Protein classes which had a higher connectivity in the Omnipath network such as transcription factors and protein kinases were recovered more often with SigNet (37% and 23%, respectively), while those with lower connectivity such as hydrolases and other enzymes had a much lower recovery (1% and 0.7%, respectively). In the case of the MetaBase™ network, we found that nuclear receptors were recovered frequently (22%) despite their relatively low connectivity compared to protein kinases and other cytosolic proteins. This could be due to the fact that nuclear receptors are just upstream from transcription factors[53] hence such targets are recovered more easily from transcriptomics data. The findings were consistent with the observed correlations in Figure 5, in that the protein class recovery was more dictated by node connectivity with the Omnipath network compared to the MetaBase™ network.

We next sought to understand how the recovery of signalling pathways related to protein class and function. To this end, we plotted the distribution of compound target-associated pathway enrichment significance for compounds targeting proteins in different classes (Figure 7). We chose to focus on SigNet and CARNIVAL with Omnipath because both showed a high performance, SigNet in terms of statistical significance, and CARNIVAL in terms of recovering informative pathways. In general, we found the highest performance for compounds targeting protein kinases and ligand-gated ion channels, which are both key mediators of cellular signalling – protein kinases transmit cellular signals through phosphorylation, and ligand-gated ion channels function to receive and transmit signals. The worst performance was seen for compounds targeting nicotinic acetylcholine and monoamine receptors; these receptors modulate signalling in the CNS [54, 55], and were additionally not expressed in high levels in the breast-cancer and prostate-cancer cell lines used (Supplementary Figure 7), hence we propose that the biological context of the cell-lines used influenced the results. We hypothesise that gene expression data measured in biological models derived from the CNS would lead to a higher recovery of such signalling pathways. We note that this particular analysis was complicated by the fact that compounds can target proteins from multiple classes.

Overall, the results of this section (Figures 5-7) show that the performance of Causal Reasoning algorithms for recovering compound targets, and compound-associated pathways, is not equal across protein classes. The connectivity of targets on the prior knowledge network were shown to heavily impact their direct recovery by the algorithms, with algorithms which correct for uneven connectivity (CausalR) showing less of an association between target degree and its successful recovery. The biological role of the considered targets was also reflected in the results, with protein kinases recovered most successfully both in terms of direct target recovery, and the recovery of relevant pathways. A potential way to mitigate the connectivity bias is for random simulation studies be carried out to identify which network nodes may be recovered by chance, an approach which has been used previously[41]. We note that one argument against this is that well-connected nodes in networks are well-studied, and have found to be essential proteins with key roles in diseases[51], and correcting against them could lead to discarding potential true positives, so that the bias described in this section is at least to an extent also desired and useful for elucidating compound targets.

Case Study – Understanding the Mechanism of Action of Chloroquine

To demonstrate the utility of causal reasoning algorithms to recover compound mechanism of action, we now present a case study for the compound Chloroquine. We chose to use CARNIVAL with the Omnipath network for this case study due to the observed good performance for recovering informative compound target-associated pathways, and the fact that the algorithm is implemented in a free, open-source R package. Briefly, we took the output for Chloroquine (CQ) generated as described in *Materials and Methods*, retaining network nodes appearing in 6 or more networks across all 12 outputs (generated with the two cell lines, two platforms, and three input gene-sets used in this study), herein named *consensus nodes*. We then used the consensus nodes to performed pathway over-representation analysis with the ReactomePA package, using all Omnipath network nodes as the universe. The top 20 most significantly enriched (BH-adjusted p-value) pathways, with their relevance to the MoA of Chloroquine (if applicable) are outlined in Table 4. The full pathway enrichment results, as well as the node counts can be found in Additional File 2. The pathways observed in Table 4 encompass literature-confirmed specific mechanisms of Chloroquine for various indications, including inflammatory mediation (MAP kinase activity, AP-1 activation, interleukin signalling, MyD88 cascade), anti-cancer action (apoptosis, mitotic cell-cycle, PKC signalling), other processes such as NMDAR-LTD, RUNX2 expression, and cellular senescence. For some pathways, namely “PIP3 activates AKT signaling” and “MyD88:MAL(TIRAP) cascade initiated on plasma membrane”, there were no literature evidence, and there were also some very general pathways enriched (“Developmental Biology”, “Diseases of signal transduction”, “Intracellular signaling by second messengers” and “Disease”). Though TLR9 signalling, a known MoA of CQ, did not appear in the top 20 by statistical significance, the “Toll Like Receptor 9 (TLR9) Cascade” was enriched with an adjusted p-value of 2.48E-05 and an Odds Ratio of 9.47 (Additional File 2).

Table 4

Reactome pathway enrichment results for consensus nodes (present in >=6 outputs across all combinations of platform, gene set and cell line) derived from CARNIVAL analysis of Chloroquine-induced gene expression using the Omnipath network. The top 20 most significantly enriched pathways (BH-adjusted p-value) are shown here with their corresponding Odd's Ratio (OR).

Pathway Name	p. Adjust	OR	MoA Relevance
Developmental Biology	$1.09x10^{-06}$	4.70	(General pathway)
MAPK targets/ Nuclear events mediated by MAP kinases	$8.07x10^{-06}$	27.45	CQ interferes with the activation of ERK-MAP kinase proteins to regulate TNF transcription for anti-inflammatory effects[56]
Activation of the AP-1 family of transcription factors	$6.13x10^{-05}$	54.91	CQ shows immunomodulatory effects in T-cells through activation of AP-1 [57]
Intrinsic Pathway for Apoptosis	$6.32x10^{-05}$	17.16	CQ was found to induce intrinsic apoptosis in cancer cells, both alone and through synergistic effects with other treatments[58–60]
Signaling by Interleukins	$6.23x10^{-05}$	4.87	CQ modulates interleukin release, promoting Th17 cell inflammation[61]
MAP kinase activation	$1.64x10^{-04}$	13.73	CQ interferes with the activation of ERK-MAP kinase proteins to regulate TNF transcription for anti-inflammatory effects[56]
Diseases of signal transduction	$1.64x10^{-04}$	5.56	(General pathway)
Transcriptional regulation by RUNX2	$1.64x10^{-04}$	13.50	Inhibition of lysosome function by CQ in vascular smooth muscle cells significantly enhanced RUNX2 expression[62]
Interleukin-17 signaling	$1.76x10^{-04}$	13.07	CQ-treated Langerhans-like cells promoted IL-17 secretion by T cells[61]
Intracellular signaling by second messengers	$1.96x10^{-04}$	6.21	(General pathway)
Gastrin-CREB signalling pathway via PKC and MAPK	$2.34x10^{-04}$	30.51	CQ activates p38 MAPK and stimulates PKC translocation in glioma cells, and Gastrin-releasing peptide has been found to mediate the CQ itch response[63, 64]
Cell Cycle, Mitotic	$2.50x10^{-04}$	5.07	CQ treatment shows anti-cancer effects through G(2)/M (mitotic) phase arrest in a human breast cancer cell line, and potentiates the effect of anti-mitotic drugs in resistant cancer cells[65],[66]
Senescence-Associated Secretory Phenotype (SASP)	$2.68x10^{-04}$	11.44	CQ has been shown to significantly inhibit the widely-used biomarker of cellular senescence, (beta)-galactosidase, in endothelial cells[67]

Pathway Name	p. Adjust	OR	MoA Relevance
CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signalling	$2.68x10^{-04}$	27.45	A study of CQ-induced autophagy inhibition found an induction of NMDAR-LTD (NMDA receptor-dependent long-term depression) [68]
Toll Like Receptor 10 (TLR10) Cascade	$2.68x10^{-04}$	10.84	No literature evidence for TLR10 and TLR5, but CQ has been shown to inhibit TLR9 signalling[69]
Toll Like Receptor 5 (TLR5) Cascade	$2.68x10^{-04}$	10.84	
MyD88 cascade initiated on plasma membrane	$2.68x10^{-04}$	10.84	CQ blocks MyD88 signaling by decrease in the levels of downstream signalling molecules[70]
Disease	$2.68x10^{-04}$	3.79	(General pathway)
PIP3 activates AKT signaling	$2.77x10^{-04}$	6.50	No literature evidence
MyD88:MAL(TIRAP) cascade initiated on plasma membrane	$2.77x10^{-04}$	10.17	No literature evidence for this specific pathway

This case study has hence demonstrated the use of causal reasoning algorithms with gene expression data for informing of specific mechanistic pathways for drug action, including off-target interactions such as CQ-induced itch response. We note that none of the known targets of CQ appeared in the consensus node list (Additional File 2), but we argue that the knowledge of targets can be derived from the pathways, and complementary approaches such as structure-based target prediction can be carried out to hypothesise direct target engagement.

Limitations of this study

While we aimed for a comprehensive parameter exploration and benchmarking of causal reasoning algorithms in this work, it still has some limitations as well. Firstly, we were limited by the annotations available for the compounds used, and chemical space (and mode of action) coverage in this set in the first place. This has profound implications for our work (and indeed, in any work where a 'ground truth' must be set for the mode of action of compounds): Different areas of chemical and mode of action space behave differently, and assuming unavailable data as inactive punishes 'false positives', which may very well be novel true positives which are just not annotated as such. Furthermore, the causal reasoning algorithms additionally infer node directionality (i.e., whether the recovered signalling proteins activated or inhibited), which we did not consider when benchmarking the results as it was not possible to obtain consistent and complete functional pharmacology information about the compound-target interactions (i.e., are they activated or inhibited upon pharmacological modulation).

Furthermore, the cell lines used in the gene expression experiments considered (MCF7 and PC3) are quite similar in terms of baseline gene expression (Supplementary Figure 7), which is why we potentially did

not see any significant difference in performance when using data derived from either cell line, but these are the only cell lines available in the original CMap. As LINCS provides data derived from 99 cell types[9], it would be interesting to investigate these other cell lines and relate these to the applicability domain of the methodologies – for example, non-cancer cell lines from a variety of tissues such as HA1E (normal kidney) or CNS cells such as NPC/NEU. It is also important to note that cell lines are *in vitro* models which cannot necessarily recapitulate *in vivo* processes, and the limitations of using cell lines in general has been extensively discussed[71, 72].

Conclusions

In this work we performed a comprehensive benchmarking of the SigNet, CausalR and CARNIVAL causal reasoning algorithms to recover compound mechanism of action from L1000 and CMap transcriptomics data, measured in the MCF7 and PC3 cell lines, and using four different prior knowledge networks. By considering two evaluation metrics, on the direct target level and the pathway level, and performing a full-factorial fixed effects analysis, we were able to identify the factors which had the most significant influence on MoA recovery.

In terms of the performance of recovering direct targets and compound-associated pathways, we found that the choice of network and algorithm were the two most significant individual factors according to the ANOVA models ($p < 2x10^{-16}$, Table 3), additionally the Network:Algorithm interaction effect was significant ($p < 2x10^{-16}$, Table 3), indicating that networks behave rather differently with different prior networks. Additionally, for the recovery of compound-associated pathways, the Algorithm:Gene set interaction effect was significant ($p = 0.001$, Table 3). The other individual factors and interaction effects, namely the choice of platform (LINCS L1000 or CMap microarray), cell line (MCF7 or PC3), and gene set (landmark, landmark and best inferred, or all genes) were not found to be statistically significant ($p > 0.05$, Supplementary Tables 1 & 2).

We hence performed post-hoc least square means analysis on the interaction effects (Network:Algorithm for both evaluation metrics, and Algorithm:Gene set for the pathway recovery metric). We found that the Algorithm:Gene set effect was relatively small so as not to be practically relevant; a small but statistically significant increase in performance was observed with CARNIVAL using only landmark genes (-log10(p-value) least square mean of 4.16 for landmark genes vs. 3.29 for all genes, Supplementary Figure 2). Notably, we found that algorithms behaved rather differently with different prior networks (Figures 3 and 4). In particular, the larger MetaBase™ networks were found to be more suitable with the CausalR scoring function to discover *direct* compound targets – on the other hand, SigNet with the smaller Omnipath network achieved the highest performance for direct target recovery. By calculating the enrichment factor of true targets in the top 5% of the ranked lists (CausalR and SigNet), we found that SigNet was better able to recall true targets early with Omnipath compared to MetaBase™, and vice versa for CausalR (Supplementary Figure 3). CARNIVAL is on the other hand less appropriate for discovering direct targets – however, this algorithm was able to recapitulate networks representative of specific and informative *pathways* encompassing actual compound targets. The CausalR ranked table performed the worst at all

metrics; we hence suggest using the ScanR function to derive consensus drivers from which a subnetwork can be reconstructed (CausalR subnetwork), an approach which performed moderately across both evaluation metrics, or, if a ranked table is required, to use SigNet instead, which showed the highest performance of all metrics on the datasets used here. From this analysis we can thus conclude that the right combination of algorithm and network needs to be employed when using causal reasoning for mode of action analysis.

Target recovery heavily depends on target connectivity (number of edges) in a given network. We found that the CausalR ranked table, which corrects for degree bias, had the lowest concordance between connectivity and recovery. This has two effects: On the one hand, this penalises hub nodes, which often represent disease targets (52) (and this lack of bias may hence also be related to the relatively low performance in terms of direct target recovery). On the other hand, for atypical (or incompletely annotated targets) this type of behaviour may be rather beneficial. Furthermore, we found that the biological role of the proteins influenced their successful recovery: transcription factors and nuclear receptors were most often recovered as direct targets, and protein kinases and mediators of cellular signalling on the pathway level. These findings are notable as they show that network and method bias, as well as target class, make big differences with respect to the performance of the algorithms employed here for mode of action analysis.

Overall, we were in this work hence able to explore performance, and factors that influence performance, when using gene expression data in combination with causal reasoning algorithms for mechanism of action analysis, which provides guidelines for their use by researchers in this field in the future.

Abbreviations

MoA
Mechanism of Action
Differentially Expressed Gene(s)
DEG(s)
CMap
Connectivity Map
PPI
Protein-Protein Interaction
PKN
Prior Knowledge Network
LM
Landmark
TF
Transcription Factor
CQ
Chloroquine

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The transcriptomics datasets used in this study are available in the GEO (https://www.ncbi.nlm.nih.gov/geo), accession numbers *GSE92742* and *GSE70138* and https://portals.broadinstitute.org/cmap. The Omnipath network used in this study is available at https://omnipathdb.org/. The MetaBase™ networks used in this study are available from Clarivate™, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Competing interests

The authors declare that they have no competing interests

Funding

L.H-G thanks the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/M011194/1] and Eli Lilly and Company for funding.

Authors Contributions

L.H-G performed and interpreted the data analyses. DC, EL, DE, HB and AB supervised the study. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the University of Cambridge Stats Clinic for helpful discussion and guidance with the statistical analyses performed in this study.

References

1. Iwata M, Sawada R, Iwata H, Kotera M, Yamanishi Y. Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. Sci Rep. 2017;7.

2. Trapotsi M-A, Hosseini-Gerami L, Bender A. Computational analyses of mechanism of action (MoA): data, methods and integration. RSC Chem Biol. 2021. https://doi.org/10.1039/D1CB00069A.
3. Casas AI, Hassan AA, Larsen SJ, Gomez-Rangel V, Elbatreek M, Kleikers PWM, et al. From single drug targets to synergistic network pharmacology in ischemic stroke. Proc Natl Acad Sci. 2019;116:7129–36.
4. Mechanism matters. Nat Med. 2010;16:347–347.
5. Berger SI, Iyengar R. Role of systems pharmacology in understanding drug adverse events. Wiley Interdiscip Rev Syst Biol Med. 2011;3:129–35.
6. Watkins J, Marsh A, Taylor PC, Singer DR. Personalized medicine: the impact on chemistry. Ther Deliv. 2010;1:651–65.
7. De Savi C, Hughes DL, Kvaerno L. Quest for a COVID-19 Cure by Repurposing Small-Molecule Drugs: Mechanism of Action, Clinical Development, Synthesis at Scale, and Outlook for Supply. Org Process Res Dev. 2020;24:940–76.
8. Whittaker PA. The role of bioinformatics in target validation. Drug Discov Today Technol. 2004;1:125–33.
9. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171:1437-1452.e17.
10. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science. 2006;313:1929.
11. Catlett NL, Bargnesi AJ, Ungerer S, Seagaran T, Ladd W, Elliston KO, et al. Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. BMC Bioinformatics. 2013;14:340.
12. Gao S, Han L, Luo D, Liu G, Xiao Z, Shan G, et al. Modeling drug mechanism of action with large scale gene-expression profiles using GPAR, an artificial intelligence platform. BMC Bioinformatics. 2021;22:17.
13. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26:511–24.
14. Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. Drug Discov Today. 2021;26:1040–52.
15. Sherman-Baust CA, Becker KG, Wood III WH, Zhang Y, Morin PJ. Gene expression and pathway analysis of ovarian cancer cells selected for resistance to cisplatin, paclitaxel, or doxorubicin. J Ovarian Res. 2011;4:21.
16. Szalai B, Saez-Rodriguez J. Why do pathway methods work better than they should? preprint. Systems Biology; 2020.
17. Uings IJ, Farrow SN. Cell receptors and cell signalling. Mol Pathol. 2000;53:295–9.

18. Enayetallah AE, Ziemek D, Leininger MT, Randhawa R, Yang J, Manion TB, et al. Modeling the Mechanism of Action of a DGAT1 Inhibitor Using a Causal Reasoning Platform. PLOS ONE. 2011;6:e27009.
19. Kumar R, Blakemore SJ, Ellis CE, Petricoin EF, Pratt D, Macoritto M, et al. Causal reasoning identifies mechanisms of sensitivity for a novel AKT kinase inhibitor, GSK690693. BMC Genomics. 2010;11:419.
20. Hill A, Gleim S, Kiefer F, Sigoillot F, Loureiro J, Jenkins J, et al. Benchmarking network algorithms for contextualizing genes of interest. PLOS Comput Biol. 2019;15:e1007403.
21. Jaeger S, Min J, Nigsch F, Camargo M, Hutz J, Cornett A, et al. Causal Network Models for Predicting Compound Targets and Driving Pathways in Cancer. J Biomol Screen. 2014;19:791–802.
22. Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, et al. Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. Cell Syst. 2018;6:484-495.e5.
23. Pabon NA, Xia Y, Estabrooks SK, Ye Z, Herbrand AK, Süß E, et al. Predicting protein targets for drug-like compounds using transcriptomics. PLOS Comput Biol. 2018;14:e1006651.
24. Verbist B, Klambauer G, Vervoort L, Talloen W, Shkedy Z, Thas O, et al. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project. Drug Discov Today. 2015;20:505–13.
25. Zhang W, Bai Y, Wang Y, Xiao W. Polypharmacology in Drug Discovery: A Review from Systems Pharmacology Perspective. Curr Pharm Des. 2016;22:3171–81.
26. Raser JM, O'Shea EK. Noise in Gene Expression: Origins, Consequences, and Control. Science. 2005;309:2010–3.
27. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci. 2001;98:4569–74.
28. Trapotsi M-A, Mervin LH, Afzal AM, Sturm N, Engkvist O, Barrett IP, et al. Comparison of Chemical Structure and Cell Morphology Information for Multitask Bioactivity Predictions. J Chem Inf Model. 2021;61:1444–56.
29. Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J, Saez-Rodriguez J. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. Npj Syst Biol Appl. 2019;5:1–10.
30. Bradley G, Barrett SJ. CausalR: extracting mechanistic sense from genome scale data. Bioinformatics. 2017;33:3670–2.
31. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40 Database issue:D1100–7.
32. Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, et al. The Drug Repurposing Hub: a next-generation drug library and information resource. Nat Med. 2017;23:405–8.
33. Enache OM, Lahr DL, Natoli TE, Litichevskiy L, Wadden D, Flynn C, et al. The GCTx format and cmap{Py, R, M} packages: resources for the optimized storage and integrated traversal of dense

matrices of data and annotations. bioRxiv. 2017;:227041.
34. Girke T. tgirke/longevityTools. R. 2019.
35. Hubbell E, Liu W-M, Mei R. Robust estimators for expression analysis. Bioinformatics. 2002;18:1585–92.
36. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47–e47.
37. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods. 2016;13:966–7.
38. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. 2019.
39. Garcia-Alonso L, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. bioRxiv. 2018;:337915.
40. Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun. 2018;9:20.
41. Fortelny N, Bock C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. Genome Biol. 2020;21:190.
42. Iskar M, Campillos M, Kuhn M, Jensen LJ, van Noort V, Bork P. Drug-Induced Regulation of Target Expression. PLoS Comput Biol. 2010;6:e1000925.
43. Yabuuchi H. enrichvs: Enrichment assessment of virtual screening approaches. 2011.
44. Truchon J-F, Bayly CI. Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem. J Chem Inf Model. 2007;47:488–508.
45. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst. 2016;12:477–9.
46. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57:289–300.
47. Neo4j Documentation. Neo4j Graph Database Platform. https://neo4j.com/docs/. Accessed 22 Sep 2020.
48. CRAN - Package neo4r. https://cran.r-project.org/web/packages/neo4r/index.html. Accessed 22 Sep 2020.
49. Searle SR, Speed FM, Milliken GA. Population Marginal Means in the Linear Model: An Alternative to Least Squares Means. Am Stat. 1980;34:216–21.
50. Picart-Armada S, Barrett SJ, Willé DR, Perera-Lluna A, Gutteridge A, Dessailly BH. Benchmarking network propagation methods for disease gene identification. PLOS Comput Biol. 2019;15:e1007276.
51. Schaefer MH, Serrano L, Andrade-Navarro MA. Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. Front Genet. 2015;6.
52. Kotlyar M, Fortney K, Jurisica I. Network-based characterization of drug-regulated genes, drug targets, and toxicity. Methods. 2012;57:499–507.

53. Sever R, Glass CK. Signaling by Nuclear Receptors. Cold Spring Harb Perspect Biol. 2013;5.
54. Dani JA. Overview of nicotinic receptors and their roles in the central nervous system. Biol Psychiatry. 2001;49:166–74.
55. Stemmer K, Müller TD, DiMarchi RD, Pfluger PT, Tschöp MH. CNS-targeting pharmacological interventions for the metabolic syndrome. J Clin Invest. 2019;129:4058–71.
56. Weber SM, Chen J-M, Levitz SM. Inhibition of Mitogen-Activated Protein Kinase Signaling by Chloroquine. J Immunol. 2002;168:5303–9.
57. Schmidt RLJ, Jutz S, Goldhahn K, Witzeneder N, Gerner MC, Trapin D, et al. Chloroquine inhibits human CD4+ T-cell activation by AP-1 signaling modulation. Sci Rep. 2017;7:42191.
58. Jia B, Xue Y, Yan X, Li J, Wu Y, Guo R, et al. Autophagy inhibitor chloroquine induces apoptosis of cholangiocarcinoma cells via endoplasmic reticulum stress. Oncol Lett. 2018;16:3509–16.
59. Monma H, Iida Y, Moritani T, Okimoto T, Tanino R, Tajima Y, et al. Chloroquine augments TRAIL-induced apoptosis and induces G2/M phase arrest in human pancreatic cancer cells. PLOS ONE. 2018;13:e0193990.
60. Ye H, Chen M, Cao F, Huang H, Zhan R, Zheng X. Chloroquine, an autophagy inhibitor, potentiates the radiosensitivity of glioma initiating cells by inhibiting autophagy and activating apoptosis. BMC Neurol. 2016;16:178.
61. Said A, Bock S, Lajqi T, Müller G, Weindl G. Chloroquine promotes IL-17 production by CD4+ T cells via p38-dependent IL-23 release by monocyte-derived Langerhans-like cells. J Immunol Baltim Md 1950. 2014;193:6135–43.
62. Cai Y, Wang X-L, Flores AM, Lin T, Guzman RJ. Inhibition of endo-lysosomal function exacerbates vascular calcification. Sci Rep. 2018;8:3377.
63. Chen T-H, Chang P-C, Chang M-C, Lin Y-F, Lee H-M. Chloroquine induces the expression of inducible nitric oxide synthase in C6 glioma cells. Pharmacol Res. 2005;51:329–36.
64. Akiyama T, Tominaga M, Takamori K, Carstens MI, Carstens E. Roles of glutamate, substance P and gastrin releasing peptide as spinal neurotransmitters of histaminergic and non-histaminergic itch. Pain. 2014;155:80–92.
65. Jiang P, Zhao Y, Shi W, Deng X, Xie G, Mao Y, et al. Cell growth inhibition, G2/M cell cycle arrest, and apoptosis induced by chloroquine in human breast cancer cell line Bcap-37. Cell Physiol Biochem Int J Exp Cell Physiol Biochem Pharmacol. 2008;22:431–40.
66. Choi A-R, Kim J-H, Woo YH, Kim HS, Yoon S. Anti-malarial Drugs Primaquine and Chloroquine Have Different Sensitization Effects with Anti-mitotic Drugs in Resistant Cancer Cells. Anticancer Res. 2016;36:1641–8.
67. Kurz DJ, Decary S, Hong Y, Erusalimsky JD. Senescence-associated (beta)-galactosidase reflects an increase in lysosomal mass during replicative ageing of human endothelial cells. J Cell Sci. 2000;113 (Pt 20):3613–22.

68. Shen H, Zhu H, Panja D, Gu Q, Li Z. Autophagy controls the induction and developmental decline of NMDAR-LTD through endocytic recycling. Nat Commun. 2020;11:2979.
69. Yasuda H, Leelahavanichkul A, Tsunoda S, Dear JW, Takahashi Y, Ito S, et al. Chloroquine and inhibition of Toll-like receptor 9 protect from sepsis-induced acute kidney injury. Am J Physiol Renal Physiol. 2008;294:F1050–8.
70. Al-Bari MAA. Targeting endosomal acidification by chloroquine analogs as a promising strategy for the treatment of emerging viral diseases. Pharmacol Res Perspect. 2017;5:e00293.
71. Hartung T, Daston G. Are In Vitro Tests Suitable for Regulatory Use? Toxicol Sci. 2009;111:233–7.
72. Hartung T. Perspectives on In Vitro to In Vivo Extrapolations. Appl Vitro Toxicol. 2018;4:305–16.

Figures

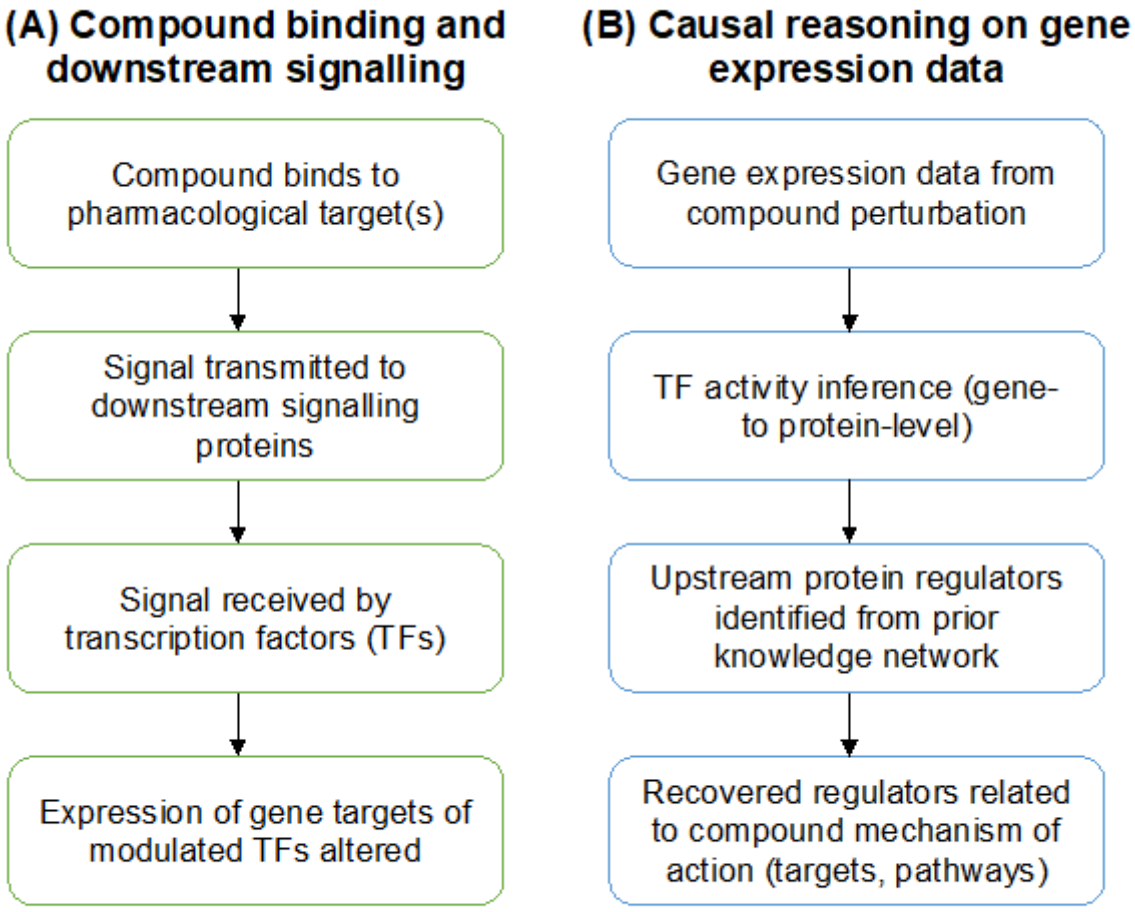

Figure 1

Biological background behind causal reasoning approaches, simplified for illustration purposes. (A) Compounds bind to their protein target(s), which elicits a cellular signal, further propagated in the cell by

signal transducer proteins, and leading to the regulation of effector proteins such as transcription factors which subsequently modulate gene expression. (B) Compound-perturbed gene expression data is leveraged to infer upstream transcription factor activity, and a network of protein-protein interactions is used to infer signalling proteins which would lead to the observed (measured) transcriptional changes. Such key driver proteins lend themselves to testable hypotheses that can then be related to the compound’s MoA in terms of targets or pathways, and validated experimentally.

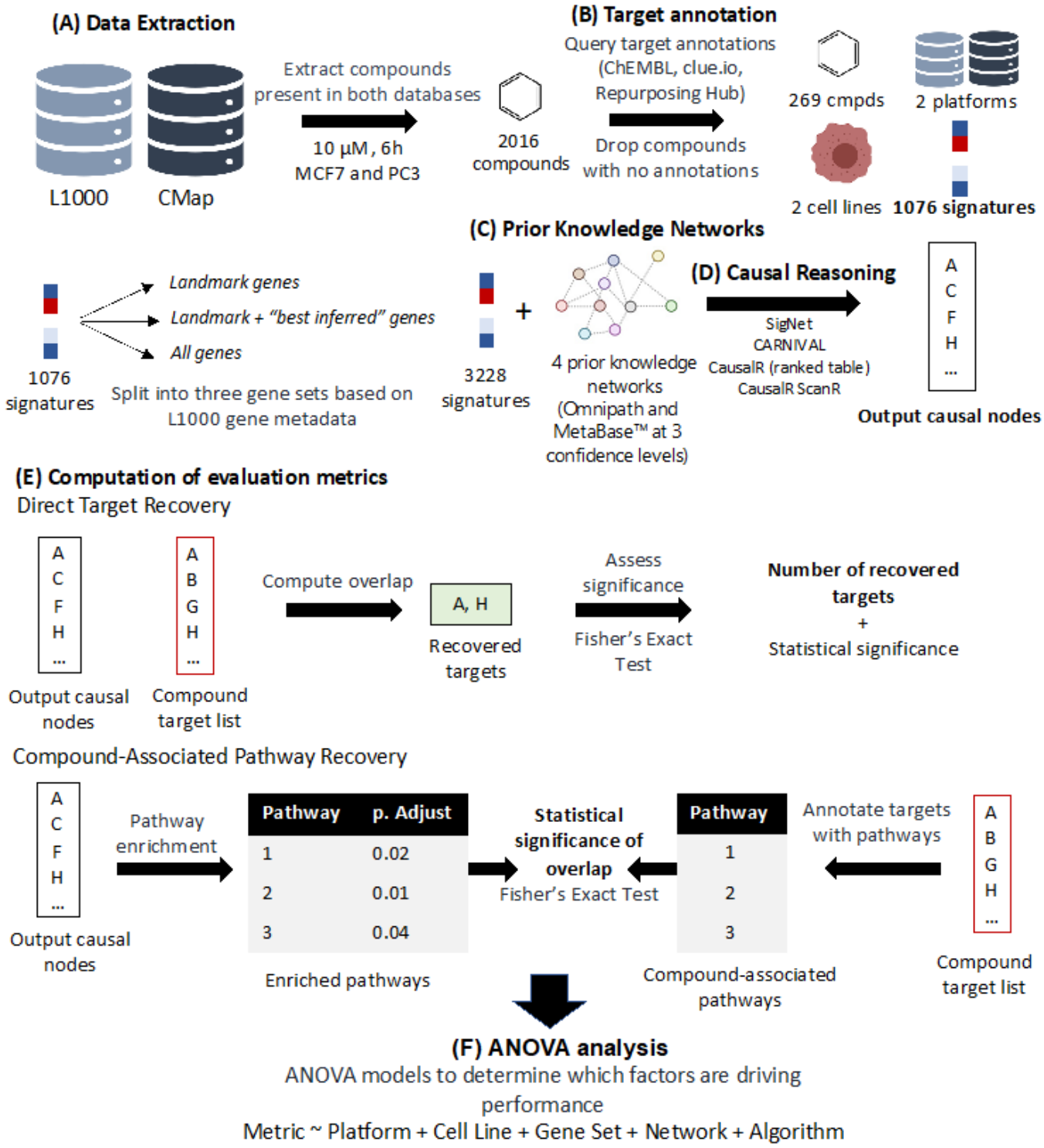

Figure 2

Workflow for the current study, involving extraction of data from L1000 and CMap measured at 10µM and 6h, in the MCF7 and PC3 cell lines. Compounds were annotated with targets and signatures split into three gene sets based on their identity in the L1000 metadata. Causal reasoning was performed and evaluation metrics computed, which were then fed into an ANOVA model to determine the most important factors for driving performance.

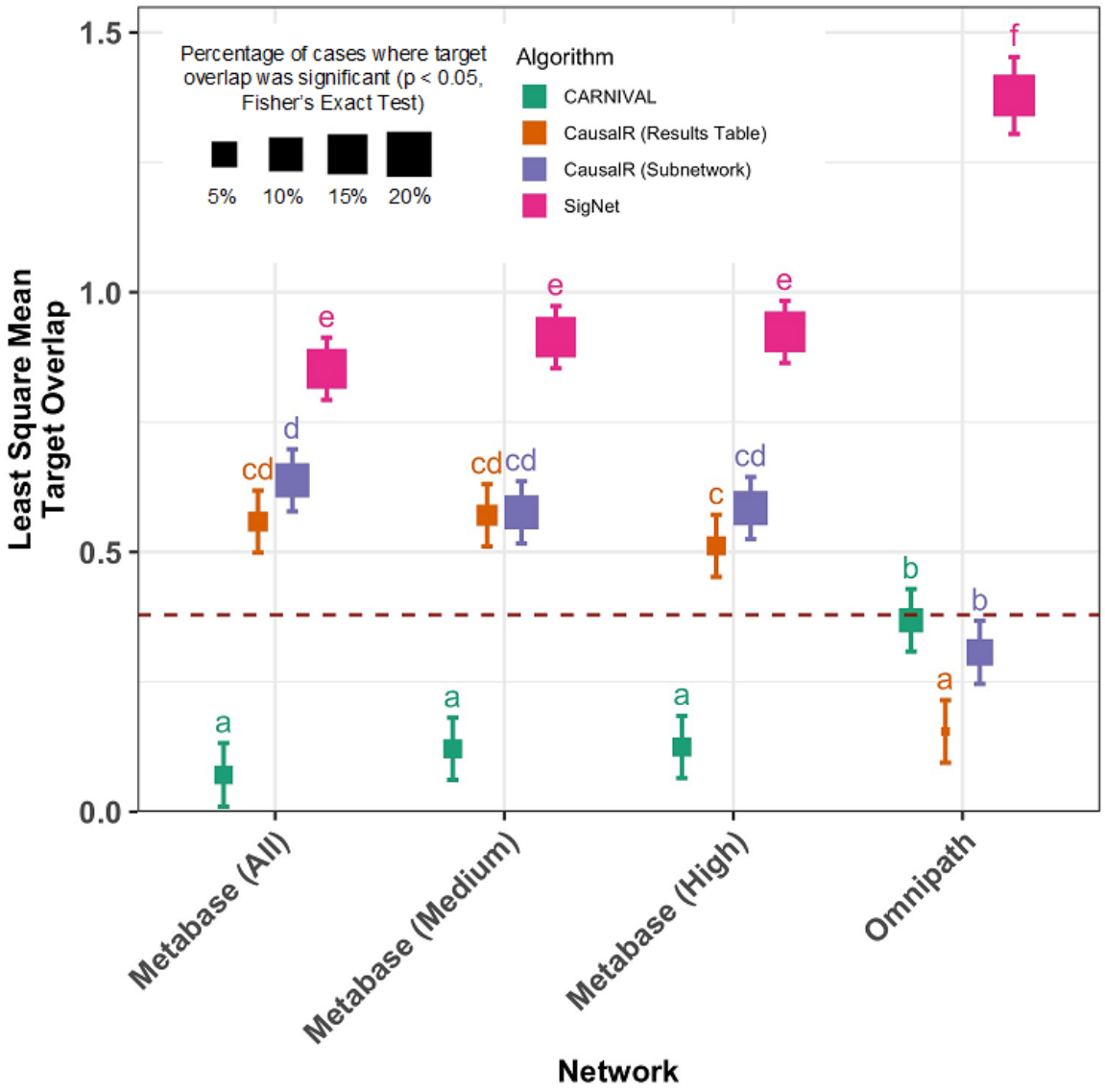

Figure 3

Interaction plot showing post-hoc least square means across all other factors for the Network:Algorithm interaction effect from the full-factorial fixed effects ANOVA model of the target recovery evaluation metric. Means sharing a letter are not significantly different according to pairwise comparisons of least square means, with Tukey adjusted p-values for multiple comparisons. Error bars indicate the least square means Tukey 95% confidence interval. The percentage of cases where the target overlap was found to be significant (p<= 0.05, Fisher's Exact Test) is encoded in point size. Baseline comparison with results from input gene expression data (direct target recovery from DEGs, averaged over platforms, gene sets, and cell lines) are indicated with a dashed red line. The SigNet algorithm with the Omnipath network achieved the highest performance for direct target recovery overall.

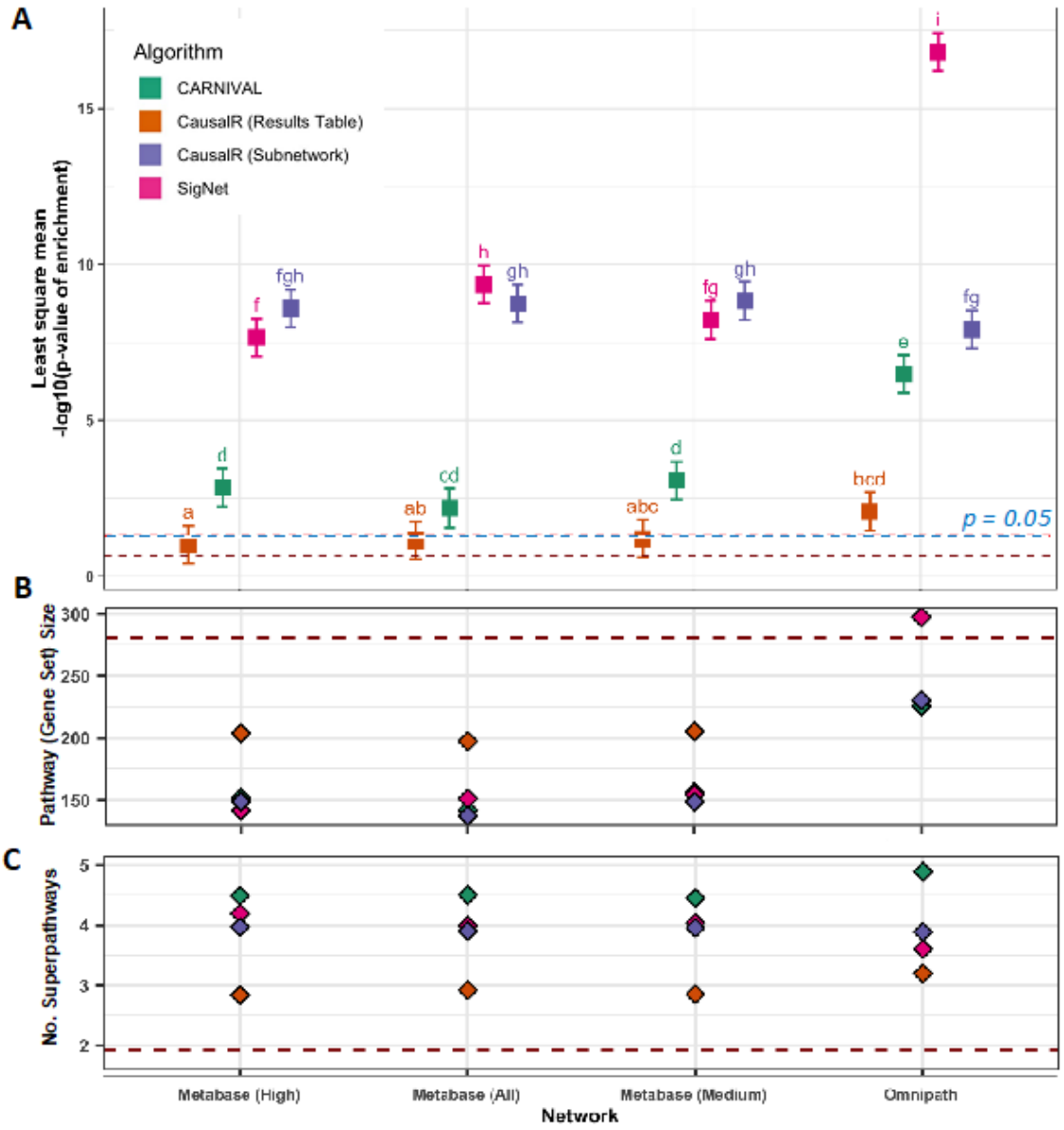

Figure 4

(A) Interaction plot showing post-hoc least square means results for the Network:Algorithm interaction effect from the full-factorial fixed effects ANOVA of compound-associated pathway recovery across all factors (over-representation -log10(p-value) of causal reasoning node-enriched pathways in the set of target annotated pathways). Means sharing a letter are not significantly different according to pairwise comparisons of least square means with Tukey adjustment for multiple comparisons. Error bars indicate the 95% confidence levels. Baseline comparison with mean results from input gene expression data (over-representation -log10(p-value) of DEG-enriched pathways in the set of target annotated pathways) is indicated with a dashed red line. Significance threshold 0.05 is indicated with a dashed blue line (~1.3)

(B) Mean Reactome pathway (gene set) sizes (number of attributed genes/proteins) for the correctly recovered pathways with baseline gene expression results indicated with a dashed red line – smaller is better (C) Mean number of superpathways within the Reactome hierarchy for the correctly recovered pathways with baseline gene expression results indicated with a dashed red line – larger is better.

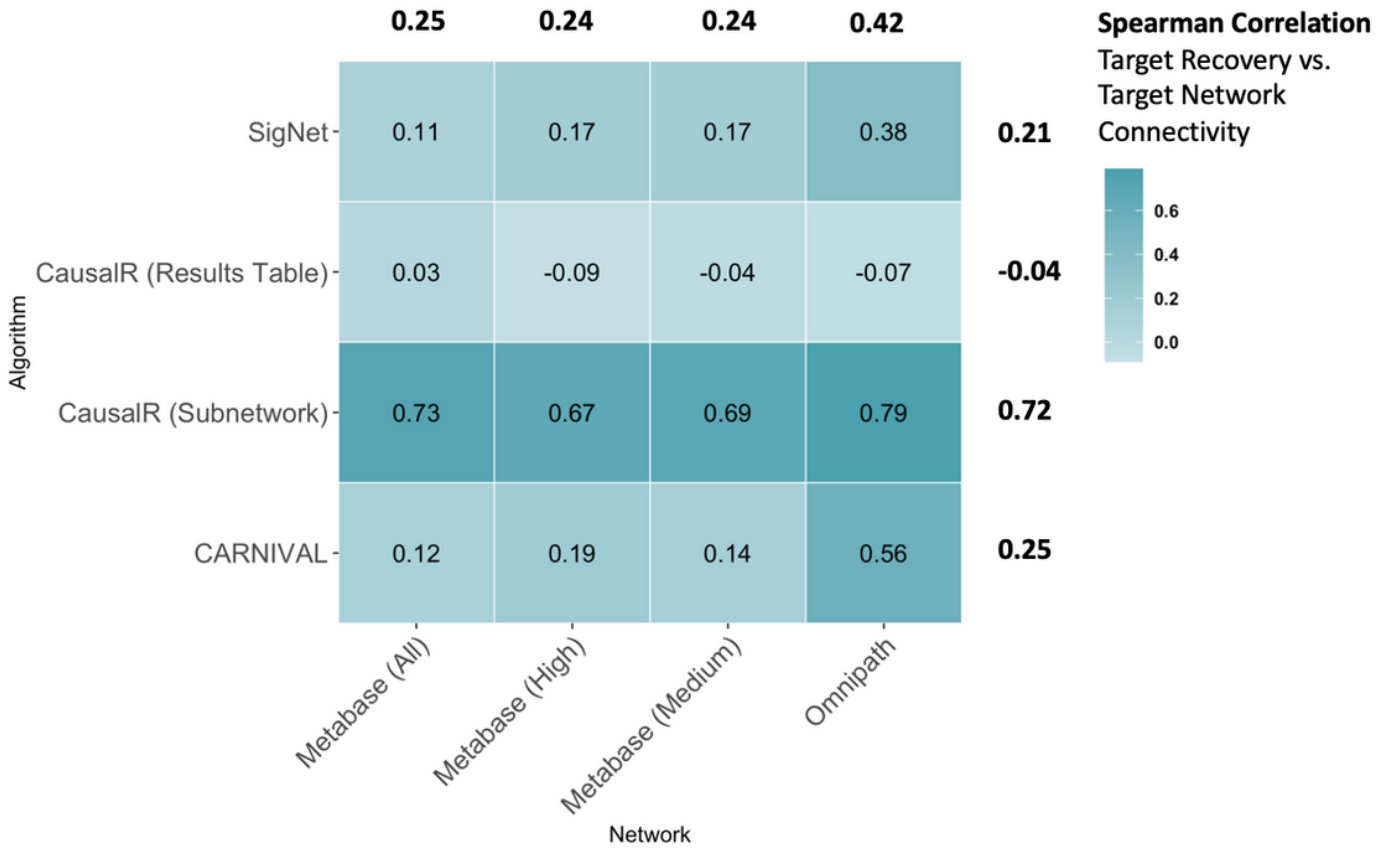

Figure 5

Heatmap showing the Spearman rank correlation of target recovery vs. target connectivity on the prior knowledge network, for each combination of network and algorithm. Mean correlation values for each network (top) and algorithm (right) are also shown.

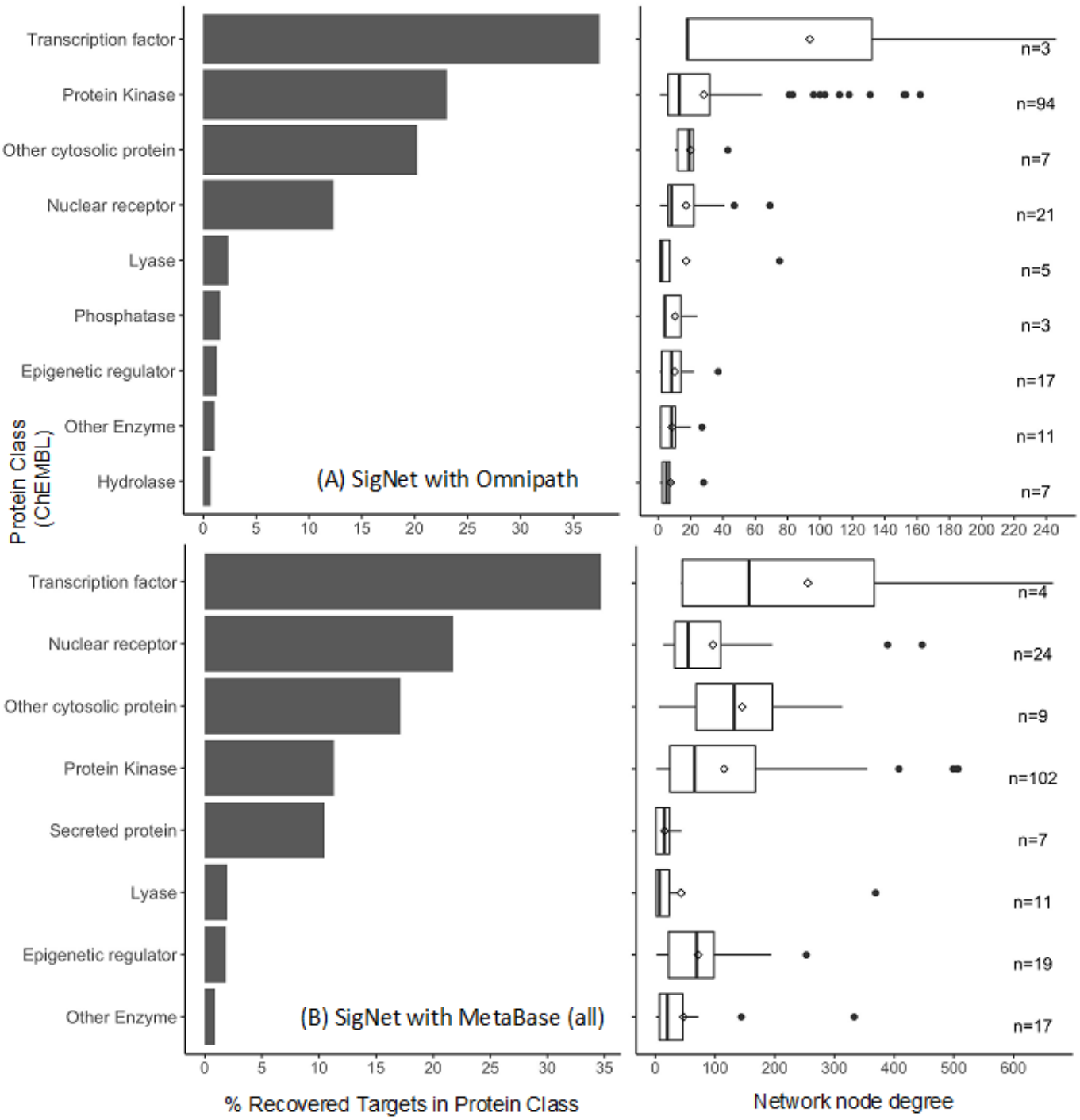

Figure 6

Protein class recovery (% of directly recovered targets in protein class) and network node connectivity for (A) SigNet with Omnipath and (B) SigNet with the full MetaBase™ network. Mean node degrees are represented by a diamond shape in the box plots, while the number of observations is labelled on the right-hand side of the plot.

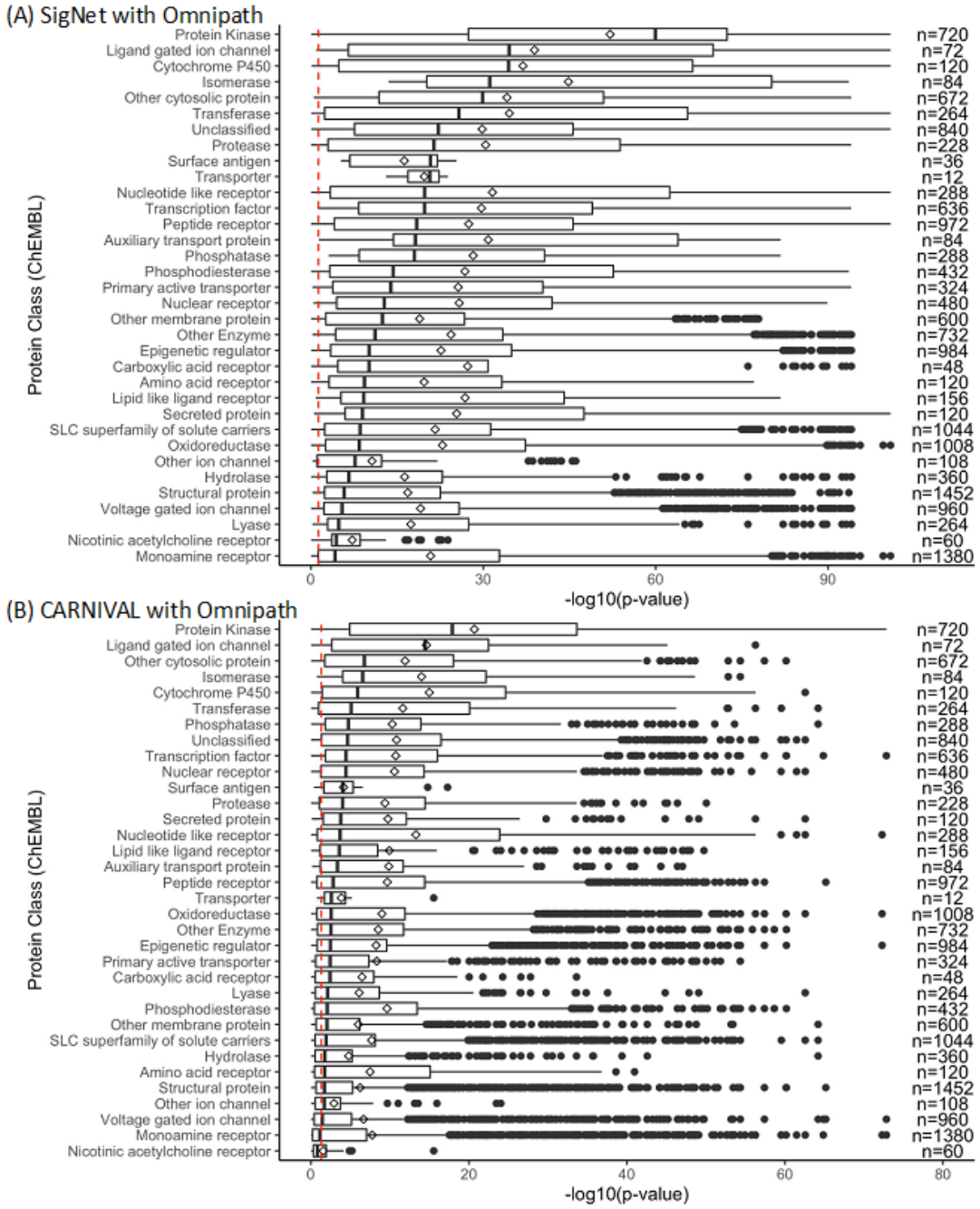

Figure 7

Distributions of the significance of compound-associated pathway enrichment based on causal nodes from (A) SigNet and Omnipath and (B) CARNIVAL and Omnipath, across all combinations of factors, separated into the protein classes targeted by each compound. Mean values are indicated by a diamond shape in the box plot. The number of instances is annotated on the right-hand side of the plot. The significance threshold p = 0.05 is indicated with a red dashed line.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.xlsx
- AdditionalFile2.xlsx
- SupplementaryInformation.docx