Inferring causal molecular networks: empirical assessment through a community-based effort

Steven M. Hill[1,*], Laura M. Heiser[2,3,4,*], Thomas Cokelaer[5], Michael Unger[6], Nicole K. Nesser[7], Daniel E. Carlin[8], Yang Zhang[9,25], Artem Sokolov[8], Evan O. Paull[8], Chris K. Wong[8], Kiley Graim[8], Adrian Bivol[8], Haizhou Wang[9,25], Fan Zhu[10], Bahman Afsari[11], Ludmila V. Danilova[11,12], Alexander V. Favorov[11,12,13], Wai Shing Lee[11], Dane Taylor[14,15], Chenyue W. Hu[16], Byron L. Long[16], David P. Noren[16], Alexander J. Bisberg[16], HPN-DREAM Consortium, Gordon B. Mills[17], Joe W. Gray[2,3,4], Michael Kellen[18], Thea Norman[18], Stephen Friend[18], Amina A. Qutub[16], Elana J. Fertig[11], Yuanfang Guan[10,19,20], Mingzhou Song[9], Joshua M. Stuart[8], Paul T. Spellman[7], Heinz Koeppl[6,25], Gustavo Stolovitzky[21,^], Julio Saez-Rodriguez[5,22,^], Sach Mukherjee[1,23,24,25,^]

1. MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, UK
2. Department of Biomedical Engineering, Oregon Health and Science University, Portland, Oregon, USA
3. Center for Spatial Systems Biomedicine, Oregon Health and Science University, Portland, Oregon, USA
4. Knight Cancer Institute, Oregon Health and Science University, Portland, Oregon, USA
5. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, UK
6. Automatic Control Laboratory and Institute of Biochemistry, ETH Zurich, Zurich, Switzerland
7. Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, Oregon, USA
8. Biomolecular Engineering, UC Santa Cruz, Santa Cruz, California, USA
9. Department of Computer Science, New Mexico State University, Las Cruces, New Mexico, USA
10. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA
11. Department of Oncology, Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Maryland, USA
12. Laboratory of Systems Biology and Computational Genetics, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia
13. Laboratory of Bioinformatics, Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, Russia
14. Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, North Carolina, USA
15. Department of Mathematics, University of North Carolina, Chapel Hill, North Carolina, USA
16. Department of Bioengineering, Rice University, Houston, Texas, USA
17. Department of Systems Biology, MD Anderson Cancer Center, Houston, Texas, USA
18. Sage Bionetworks, Seattle, Washington, USA
19. Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA
20. Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, USA
21. IBM Translational Systems Biology and Nanobiotechnology, Yorktown Heights, New York, USA

22. RWTH-Aachen University Hospital, Joint Research Centre for Computational Biomedicine (JRC-COMBINE), Aachen, Germany
23. School of Clinical Medicine, University of Cambridge, Cambridge, UK
24. German Centre for Neurodegenerative Diseases (DZNE), Bonn, Germany

25. Present addresses: SimQuest Inc, Boston, Massachusetts, USA (H.W.); Amyris Inc, Emeryville, California, USA (Y.Z.); Department of Electrical Engineering and Information Technology, Technische Universitaet Darmstadt, Darmstadt, Germany (H.K.); German Centre for Neurodegenerative Diseases, Bonn, Germany (S.M.).

* equal contributions
^ corresponding

To whom correspondence should be addressed:
Dr. Sach Mukherjee
Email: sach.mukherjee@dzne.de

Dr. Julio Saez-Rodriguez
Email: saezrodriguez@combine.rwth-aachen.de

Dr. Gustavo Stolovitzky
Email: gustavo@us.ibm.com

Abstract

Inferring molecular networks is a central challenge in computational biology. However, it has remained unclear whether causal, rather than merely correlational, relationships can be effectively inferred in complex biological settings. Here we describe the HPN-DREAM network inference challenge that focused on learning causal influences in signaling networks. We used phosphoprotein data from cancer cell lines as well as *in silico* data from a nonlinear dynamical model. Using the phosphoprotein data, we scored more than 2,000 networks submitted by challenge participants. The networks spanned 32 biological contexts and were scored in terms of causal validity with respect to unseen interventional data. A number of approaches were effective and incorporating known biology was generally advantageous. Additional sub-challenges considered time-course prediction and visualization. Our results constitute the most comprehensive assessment of causal network inference in a mammalian setting carried out to date and suggest that learning causal relationships may be feasible in complex settings such as disease states. Furthermore, our scoring approach provides a practical way to empirically assess the causal validity of inferred molecular networks.

Introduction

Molecular networks are central to biological function and the data-driven learning of regulatory connections in molecular networks has long been a key topic in computational biology[1–6]. An emerging notion is that networks describing a certain biological process, for example signal transduction or gene regulation, may depend on biological context, such as cell type, tissue type, or disease state[7,8]. This has motivated efforts to elucidate networks that are specific to such contexts[9–14]. In disease settings, networks specific to disease context could improve understanding of the underlying biology and potentially be exploited to inform rational therapeutic interventions.

In this study, we considered inference of causal molecular networks, focusing specifically on signaling downstream of receptor tyrosine kinases. We define edges in causal molecular networks ("causal edges") as directed links between nodes in which inhibition of the parent node can lead to a change in abundance of the child node (**Fig. 1a**), either by direct interaction or via unmeasured intermediate nodes (**Fig. 1b**). Such edges may be specific to biological context (**Fig. 1c**). The notion of a causal link is fundamentally distinct from a correlational one (**Fig. 1d**). Causal network inference is profoundly challenging[15,16] and many methods for inferring regulatory networks connect together correlated, or mutually dependent, nodes that may not have any causal relationship. Some approaches (e.g. causal directed acyclic graphs[17–19]) are intended to infer causal relationships, but their success can only be guaranteed under very strong assumptions[15,20] that are almost certainly violated in biological settings. This is due to many limitations – some possibly fundamental – in our ability to observe and perturb biological systems.

These observations imply that careful empirical assessment is essential to learn whether computational methods can provide causal insights in a specific biological setting of interest. Network inference methods are often assessed using data simulated from a known causal network structure (a so-called "gold-standard" network[5,17]). Such studies (and their synthetic biology counterparts[21]) are convenient and useful but at the same time limited because it is difficult to truly mimic the features of a specific biological system. Networks inferred from experimental data are often compared to the literature. However, since the goal of network inference is to learn novel regulatory relationships that could be specific to context, this is an inherently limited approach. Hypotheses generated by computational tools can be validated experimentally, but to date such assessment has been limited[9,10,19,22].

Motivated by these observations, and with the support of the Heritage Provider Network (HPN), we developed the HPN-DREAM challenge to assess the ability to learn causal networks and predict molecular time-course data. The Dialogue for Reverse Engineering

Assessment and Methods (DREAM) project[23] (http://dreamchallenges.org) has run several challenges focused on network inference[22,24–27]. Here we focused on causal signaling networks in human cancer cell lines. Protein assays were carried out using reverse-phase protein lysate arrays[28,29] (RPPA) that included functional phosphorylated proteins.

The HPN-DREAM challenge comprised three sub-challenges. **Sub-challenge 1:** Here, the task was to infer causal signaling networks using protein time-course data. To focus on networks specific to genetic and epigenetic background, the task spanned 32 different contexts, each defined by a combination of cell line and stimulus, and each with its own training and test data. The test data were used to assess the causal validity of inferred networks, as described below. A companion *in silico* data task also focused on causal networks but by design did not allow the use of known biology. **Sub-challenge 2:** Participants were tasked with predicting phosphoprotein time-course data under perturbation. The sub-challenge comprised both an experimental data task and an *in silico* data task and the same training datasets were used as in sub-challenge 1. **Sub-challenge 3**: Participants were asked to develop methods to visualize these complex, multi-dimensional datasets.

In total across all sub-challenges, the scientific community contributed 178 submissions. In the network inference sub-challenge we found that several submissions achieved statistically significant results, providing substantive evidence that causal network inference may be feasible in a complex, mammalian setting (we discuss a number of relevant caveats below). The use of pre-existing biological knowledge (e.g. from online databases) appeared to be broadly beneficial. On the other hand, “FunChisq”, a method that did not incorporate any known biology whatsoever, was not only the top performer in the *in silico* data task, but also highly ranked in the experimental data task.

Challenge data, submissions and code are made available as a community resource through the Synapse platform[30], which was used to run the challenge (https://www.synapse.org/HPN_DREAM_Network_Challenge). Additionally, see **Supplementary Notes 1-3** for descriptions of methods applied in the challenge.

Results

Training data for network inference

For the experimental data network inference task, participants were provided with RPPA phosphoprotein data from four breast cancer cell lines under eight ligand stimulus conditions. The 32 *(cell line, stimulus)* combinations each defined a biological context. Data for each context comprised time-courses for ~45 phosphoproteins (the set of phosphoproteins varied slightly between contexts; **Supplementary Table 1**). The training data included time-courses obtained under three kinase inhibitors and a control (DMSO, **Fig. 2a**; see Online Methods for details of experimental design, protocol, quality control and pre-processing). The dataset is also available in an interactive online platform (http://dream8.dibsbiotech.com) using the “Biowheel” design developed by the winning team of the visualization sub-challenge.

Participants were tasked with using the training data to learn causal networks specific to each of the 32 contexts. Networks had to comprise nodes corresponding to each phosphoprotein with directed edges between the nodes. The edges were required to have weights indicating strength of evidence in favor of each possible edge, but did not need to indicate sign (i.e. whether activating or inhibitory). For the companion *in silico* data task, participants were provided with data generated from a nonlinear differential equation model of signaling[12]. The task was designed to mirror some of the key features of the experimental setup and participants were asked to infer a single directed, weighted network (Online Methods; **Supplementary Fig. 1**). While the experimental data task tested both data-driven learning and use of known biology, the *in silico* data task focused exclusively on the former,

and for that reason node labels (i.e. protein names in the underlying model) were anonymized.

Empirical assessment of causal networks

Standard statistical assessments of goodness-of-fit or predictive ability are not suited to assessing causal network inference. This is due to the fact that an incorrect causal network can nonetheless score very well on such metrics (e.g. two nodes that are highly correlated but not causally linked (**Fig. 1d**) may predict each other well). We therefore developed a procedure that leveraged interventional data to assess the causal validity of networks submitted to the experimental data task. The key concept was to assess the extent to which causal relationships encoded in inferred networks were in agreement with test data obtained under an entirely unseen intervention (**Fig. 2a**). Specifically, for a given context *c*, we identified the set of nodes that showed salient changes under a test inhibitor (here, inhibition of mTOR) relative to DMSO control (**Fig. 2b**; Online Methods). These nodes are causally influenced by the inhibitor target (mTOR) and can be regarded as *descendants* of the target in the underlying causal network for context *c.* We denote this "gold-standard" descendant set by D_c^{GS} (**Supplementary Fig. 2**; note that D_c^{GS} may include both downstream nodes and those influenced via feedback loops within the experimental timeframe). For each submitted context-specific network, we computed a predicted set of descendants of mTOR, which we call D_c^{pred}. We then compared D_c^{GS} to D_c^{pred} to obtain an area under the receiver operating characteristic curve (AUROC) score (**Fig. 2c**) for each context *c*, resulting in a set of 32 AUROC scores for each team. These were used to rank teams within each context. An overall score was obtained by computing the mean rank across contexts (see Online Methods) and this determined the final ranking (**Fig. 2d**, **Table 1** and **Supplementary Fig. 3a**). We tested robustness of the rankings using a subsampling strategy (see Online Methods). In **Table 1** we include mean AUROC scores across the 32 contexts. Mean AUROC scores complement mean ranks by giving information on the absolute level of performance (the two metrics are highly correlated; see **Supplementary Fig. 3c**).

For the *in silico data* task, the true causal network was known (Online Methods; **Supplementary Fig. 4**) and this was used to obtain an AUROC score for each participant that determined the final rankings (**Table 1** and **Supplementary Fig. 3b**).

An alternative scoring metric to AUROC is area under the precision-recall curve (AUPR), which is often used when there is an imbalance between the number of positives and negatives in the gold-standard[31]. Some of our gold-standard datasets were imbalanced and we therefore compared rankings based on AUROC and AUPR, finding reasonable agreement (Online Methods; **Supplementary Figs. 5** and **6**).

Performance of individual teams and ensemble networks

For the experimental data network inference task, **Figure 3a** shows the 32 AUROC scores for each team. 22 teams attained significant AUROC scores (FDR < 0.05; see Online Methods) in at least 25% of the contexts and 8 teams attained significant scores in at least 50% of contexts (**Supplementary Fig. 7a**). Conversely, for 25 out of the 32 contexts, 5 or more teams attained significant AUROC scores (**Supplementary Fig. 7b**). For the *in silico* data task, the top 14 teams achieved significant AUROC scores (**Supplementary Fig. 3b**). The fact that several teams achieved significant scores with respect to causal performance metrics suggests that causal network inference may be feasible in this setting.

Scores on the experimental data and *in silico* data network inference tasks were modestly correlated ($r = 0.35$, $p = 0.011$), but better correlated when comparing only teams that did not use prior information ($r = 0.68$, $p = 0.002$; **Fig. 3b** and **Supplementary Note 4**). To identify teams that performed well across both tasks we calculated a combined ranking (average of ranks for experimental and *in silico* data tasks; **Table 1, Fig. 3b** and **Supplementary Fig.**

3d). **Table 1** shows scores and method summaries for submissions ranked highly in either task or under the combined ranking; information for all teams can be found in **Supplementary Table 2**.

To test the notion of "crowdsourcing"[22,27,32,33] for causal network inference, we combined inferred networks across all teams and assessed the resulting ensemble or aggregate submission (Online Methods; **Fig. 3c** and **Supplementary Fig. 8a**). For the experimental data task, the aggregate submission slightly outperformed the highest-ranked submission (mean AUROC of 0.80 and 0.78 respectively) and, for the *in silico* data task, it ranked within the top 5 (AUROC = 0.67). Combinations of as few as 25% of randomly chosen submissions performed well on average (mean AUROC of 0.72 and 0.64 for experimental and *in silico* data tasks respectively; **Fig. 3d** and **Supplementary Fig. 8b**).

A total of 41 of the 80 participating teams provided details about methods used (**Supplementary Note 1)** and we used this information to classify the submissions (**Fig. 3e**,**f**,**Table 1** and **Supplementary Note 5**). In line with findings from previous DREAM challenges[22,32], we observed no clear relationship between method class and performance. However, we note that the boundaries between method classes are not always well defined and that performance can be influenced by additional factors, including details of pre-processing and implementation.

Top-performing methods for causal network inference

The best-scoring method for the experimental data task, "PropheticGranger with heat diffusion prior" by Team1, used a prior network created by averaging similarity matrices obtained via simulated heat diffusion applied to links derived from the Pathway Commons database[34]. This was then coupled with an L1-penalized regression approach that considered not only past but also future time points (see **Supplementary Note 1** for a detailed description). The best scoring approach for the *in silico* data network inference task, and the most consistent performer across both data types, was the "FunChisq" method by Team7 (see **Supplementary Note 1**). This approach used a novel functional chi-square test to examine functional dependencies among the variables and did not use any biological prior information. Before applying FunChisq, the abundance of each protein was discretized by the *Ckmeans.1d.dp* method[35], with the number of discretization levels automatically selected using the Bayesian information criterion on a Gaussian mixture model.

Incorporating pre-existing biological knowledge

On average, teams that used prior biological information out-performed those that did not (**Fig. 4a**; one-sided rank-sum test, $p = 0.032$). The submission ranked second used only a prior network and did not use the protein data. However, use of a prior network did not guarantee good performance (mean AUROC scores ranged from 0.49 to 0.78 for teams using a prior network). Interestingly, the same prior network that was itself ranked second was used in both the top-performing submission and the submission ranked 43rd, the difference being the approach used to analyze the experimental data. Conversely, not using a prior network did not necessarily result in poor performance (mean AUROC scores ranged from 0.49 to 0.71 for teams not using a prior network). The top-performing teams using prior networks in the experimental data task did not perform as well in the *in silico* data task (**Fig. 3b**).

To further investigate the influence of known biology, we combined submitted prior networks to form an *aggregate prior network* (Online Methods). This outperformed the individual prior networks and had a similar score to the aggregate submission described above (mean AUROC = 0.79). We combined the aggregate prior network with each of the two top methods (PropheticGranger and FunChisq) in varying proportions (**Fig. 4b**). Combining FunChisq with the aggregate prior improved upon the aggregate prior alone (this was not the case for PropheticGranger). Finally, we considered three-way combinations of

PropheticGranger, FunChisq and the aggregate prior; the highest-scoring combination consisted of 20% PropheticGranger, 50% FunChisq and 30% aggregate prior (mean AUROC of 0.82, **Supplementary Fig. 9**). The combination weights were set by optimizing performance on the test data itself; we note that since additional test data were not available we cannot rigorously assess the combination analyses.

Context-specific performance

For the experimental data task, participants were required to infer networks specific to each context. The overall scoring metric is an average over all contexts; to gain additional insight we further investigated performance by context. **Figure 4c** shows, for each context, performance of the aggregate submission and aggregate prior, together with the top 25 AUROC scores. In line with their good overall performance, the aggregates performed well relative to individual submissions in most contexts. The aggregate prior network performed particularly well for cell line MCF7, while in BT549 it performed less well for several stimuli. This supports the notion that biological contexts differ in the extent to which they agree with known biology. The aggregate submission offered the largest improvements over the aggregate prior in settings where the latter performed less well, suggesting that combining data-driven learning with known biology may offer the most utility in non-canonical settings.

Crowdsourced context-specific signaling hypotheses

The context-specific aggregate submission networks provide crowdsourced signaling hypotheses; one such network is shown in **Figure 5a**. Comparing the aggregate submission networks with the aggregate prior network helps to highlight potentially novel edges: a list of context-specific edges with their associated scores is provided as a resource in **Supplementary Table 3**. Dimensionality reduction suggested that differences between cell lines are more prominent than between stimuli for a given cell line (**Fig. 5b**; Online Methods), in line with the notion that (epi)genetic background plays a key role in determining network architecture.

Time-course prediction sub-challenge

Here, participants were tasked with predicting phosphoprotein time-courses obtained under interventions not seen in the training data (Online Methods). We assessed predictions by direct comparison with the test data using root mean squared error (RMSE; Online Methods and **Supplementary Note 6**). In contrast to the causal network inference sub-challenge, the focus was on predictive ability rather than causal validity. Submissions are summarized in **Supplementary Table 4** and **Supplementary Note 2**. Testing robustness of team ranks gave two top-performers for the experimental data task and a single top-performer for the *in silico* data task (Online Methods;). The two top-performers for the experimental data task took different approaches. Team42 (ranked second) simply calculated averages of values in the training data. Team10 (ranked third) used a truncated singular value decomposition to estimate parameters in a regression model. This method also ranked highly for the *in silico* data task and was the most consistent performer across both data types. Team44, the top-ranked team, was not eligible to be named as a top-performer (due to an incomplete submission; **Supplementary Note 7**), but their approach also consisted of calculating averages (the good performance of averaging may be explained to some degree by a shortcoming with the RMSE metric used here, see **Supplementary Fig. 10**). Team34, the top-performer for the *in silico* data task, used a model informed by networks learned in the network inference sub-challenge. This suggests that network inference can also play a useful role in purely predictive analyses.

Visualization sub-challenge

In total, 14 teams submitted visualizations that were made available to the HPN-DREAM Consortium members who then voted for their favorite (Online Methods). The winning entry “Biowheel” is designed to enhance the visualization of time course protein data and aid in its interpretation (see **Supplementary Note 3**; http://dream8.dibsbiotech.com). The data

associated with a cell line are plotted to depict protein abundance levels by color, as in a heat map, but displayed as a ring, or wheel. Time is plotted along the radial axis and increases from the center outwards. The interactive tool provides a way to mine data by displaying data subsets in various ways.

Discussion

Inferring molecular networks remains a key open problem in computational biology. This study was motivated by the view that empirical assessment will be essential in catalyzing the development of effective methods for causal network inference. Such methods will be needed to systemically link molecular networks to the phenotypes they influence. While there are many theoretical and practical reasons why causal network inference may fail, our results, obtained via a large-scale, community effort with blinded assessment, suggest that the task may be feasible in complex mammalian settings. By "feasible" we mean reaching a performance level significantly better than chance, and this was achieved by a number of submissions, including approaches that did not use any prior information. Nevertheless, our approach and findings are subject to caveats that we discuss below.

We put forward an assessment approach that focuses on causal validity and is general enough to be potentially applicable in a variety of biological settings. However, it is important to take note of several caveats. First, the procedure relies on specificity of test inhibitors. However, if the inhibitor were highly non-specific, it would likely not be possible to achieve good results, nor for a prior network to perform well, because predictions themselves are based on assumed specificity. In addition, data suggest that the mTOR inhibitor used here does have good specificity. Second, the procedure used only one of the inhibitors for testing. Although not possible in a "live" challenge setting, as training and test data must be fixed at the outset, Hill, Nesser *et al*.[36] used a cross-validation-type scheme that iterated over inhibitors. Such an approach may provide a more comprehensive assessment and indeed the ranking of methods could change when including additional inhibitors. Third, the procedure does not distinguish between direct and indirect causal effects. Finally, all downstream targets, whether context-specific or not, were weighted equally. Metrics that emphasize context-specific effects will be an important avenue for future research and would likely shed further light on the utility of priors (that are not usually context-specific).

Several submissions used novel methods or incorporated novel adaptations to existing methods (**Supplementary Tables 2 and 4**). Notably, the best performing team for the network inference *in silico* data task developed a novel procedure (FunChisq) that also performed well on the experimental data task without use of any prior information, increasing confidence in its robustness. Indeed, the ability to make such comparisons is a key benefit of running experimental and *in silico* challenges in parallel. Although some approaches performed well on one data type only (**Fig. 3b**), the overall positive correlation between experimental and *in silico* scores is striking given that they were based on different data and assessment metrics. Teams that did not use prior information were relatively well correlated (**Fig. 3b**), suggesting that good performers among these teams on the *in silico* data task could perform competitively on experimental data if extended to incorporate known biology.

The observation that prior information alone performs well reflects the fact that much is already known about signaling in cancer cells and suggests that causal networks are not entirely "re-wired" in these cells. On the other hand, our analysis revealed contexts which deviate from known biology; such deviations are likely particularly important for understanding disease-specific dysregulation and therapeutic heterogeneity. Furthermore, it is likely that the literature - and priors derived from it - is biased towards cancer and for that reason priors may be less effective in other disease settings. We anticipate that in the future a combination of known biology with data-driven learning will be important in elucidating networks in specific disease states.

A previous DREAM challenge also focused on signaling networks in cancer[26]. However, the scoring metric was predictive rather than causal (RMSE between predicted and test data points) with a penalty related to sparseness of the inferred network. Thus, the challenge did not focus on causal validity *per se,* and indeed a network with causally incorrect edges could yield a good RMSE score. Our assessment approach shares similarities with other approaches in the literature, including Maathuis *et al.*[37] who focused on inferring networks from static observational data and Olsen *et al.*[38] who used a different scoring metric, considering predicted downstream targets in close network proximity to the inhibited node.

There are several directions that future challenges could take. The causal scoring approach proposed here could be applied in other settings, e.g. gene regulatory or metabolic networks. In the interests of open and transparent science, and to provide a community resource, participants were encouraged to make source code available. In order to facilitate post-challenge analyses, future challenges may benefit from submission of executable programs that could allow for more controlled and detailed comparisons between methods.

It remains unclear to what extent the ranking of specific methods submitted to the challenge would generalize to different data-types and biological processes. In our view, it is still too early to say whether there could emerge broadly effective "out-of-the-box" methods for causal network inference, analogous to methods used for some machine learning tasks. Due to the complexity of causal learning and the many factors that could be application-specific we recommend that at the present time network inference efforts should whenever possible include at least some interventional data and that suitable scores, such as those described in this paper, should be used to assess network inference in a causal sense. Such an assessment would test whether inference is effective in the setting of interest.

Accession codes

All data used for the challenge are available through Synapse at https://www.synapse.org/HPN_DREAM_Network_Challenge.

Acknowledgements

P.T.S., S.M., G.B.M., and J.W.G. kindly provided the experimental data for this challenge prior to publication. We are grateful to the Heritage Provider Network for their support of the DREAM Challenge. This work was supported in part by the following: National Institutes of Health, National Cancer Institute grant U54 CA 112970 (J.W.G.); Susan G. Komen Foundation SAC110012 (J.W.G.); Prospect Creek Foundation grant (J.W.G); National Institutes of Health, National Institute of General Medical Sciences Award Number 1R01GM109031 (J.M.S.); National Institutes of Health, National Cancer Institute grant 5R01CA180778 (J.M.S.); National Cancer Institute of the National Institutes of Health Award Number U54CA143869 (M.F.C.); EuroinvesXgacion program of MICINN (Spanish Ministry of Science and InnovaXon), partners of the ERASysBio+ iniXaXve supported under the EU ERA-NET Plus scheme in FP7 (SHIPREC), and by the Spanish Ministry of Science and InnovaXon (MICINN) [FEDER BIO2008-0205, FEDER BIO2011-22568, EUI2009-04018] (B.O.). Royal Society Wolfson Research Merit Award (S.M.). BMBF GANI_MED consortium grant 03IS2061A (T.K.). National Library of Medicine grants R00LM010822 (X.J.) and R01LM011663 (X.J. and R.E.N.). We would like to thank Paul Kirk for comments on the manuscript and David Henriques for input into the post-challenge analysis of the *in silico* dataset.

Author Contributions

S.M.H., L.M.H., T.C., M.U., J.W.G., P.T.S. H.K., G.S., J.S.-R., and S.M. designed the challenge. J.W.G., P.T.S., N.K.N., G.B.M., and S.M. provided experimental data for use in the challenge. M.U. and H.K. provided data for the *in silico* challenge. M.K., T.N., and S.F. developed and implemented the Synapse platform used to facilitate the challenge. S.M.H., L.M.H., and T.C. performed analysis of challenge data. S.M.H., L.M.H., T.C., M.U., H.K.,

G.S., J.S-R., and S.M. interpreted the results of the challenge. D.E.C. and Y.Z. performed analyses to compare top-performing approaches submitted to the network inference sub-challenge. D.E.C., A.S., E.O.P., C.K.W., K.G., A.B., and J.M.S. designed the top-performing approach in the experimental data network inference task. Y.Z., H.W., and M.S. designed the approach that was top-performing in the *in silico* data network inference task, and was also the most highly ranked across both the experimental and *in silico* data network inference tasks. F.Z. and Y.G. developed an algorithm that was a top-performer in the experimental data time-course prediction task, and was also the most highly ranked across both the experimental and *in silico* data time-course prediction tasks. B.A., L.V.D., A.V.F., W.S.L., D.T., and E.J.F. comprised one of the top-performing teams in the experimental data time-course prediction task. C.W.H., B.L.L., D.P.N., A.J.B., and A.A.Q. designed the Biowheel visualization tool. The HPN-DREAM Community provided predictions and Supplementary Note descriptions of the algorithms. S.M.H., L.M.H., T.C., M.U., D.E.C., Y.Z., M.S., J.M.S., H.K., G.S., J.S-R., and S.M. wrote the paper.

References

1. Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & di Bernardo, D. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **3,** 78 (2007).

2. Markowetz, F. & Spang, R. Inferring cellular networks--a review. *BMC Bioinformatics* **8,** S5 (2007).

3. Hecker, M., Lambeck, S., Toepfer, S., van Someren, E. & Guthke, R. Gene regulatory network inference: data integration in dynamic models-a review. *BioSystems* **96,** 86–103 (2009).

4. De Smet, R. & Marchal, K. Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* **8,** 717–729 (2010).

5. Marbach, D. *et al.* Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 6286–6291 (2010).

6. Maetschke, S.R., Madhamshettiwar, P.B., Davis, M.J. & Ragan, M.A. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Brief. Bioinform.* **15,** 195–211 (2014).

7. Ideker, T. & Krogan, N.J. Differential network biology. *Mol. Syst. Biol.* **8,** 565 (2012).

8. De la Fuente, A. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet.* **26,** 326–33 (2010).

9. Hill, S.M. *et al.* Bayesian inference of signaling network topology in a cancer cell line. *Bioinformatics* **28,** 2804–10 (2012).

10. Saez-Rodriguez, J. *et al.* Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer Res.* **71,** 5400–11 (2011).

11. Molinelli, E. J. *et al.* Perturbation biology: inferring signaling networks in cellular systems. *PLoS Comput. Biol.* **9,** e1003290 (2013).

12. Chen, W.W. *et al.* Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol. Syst. Biol.* **5,** 239 (2009).

13. Akbani, R. *et al.* A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat. Commun.* **5,** 3887 (2014).

14. Eduati, F., De Las Rivas, J., Di Camillo, B., Toffolo, G. & Saez-Rodriguez, J. Integrating literature-constrained and data-driven inference of signalling networks. *Bioinformatics* **28,** 2311–2317 (2012).

15. Pearl, J. *Causality: Models, Reasoning, and Inference* 2nd edn. (Cambridge University Press, 2009).

16. Freedman, D. & Humphreys, P. Are there algorithms that discover causal structure? *Synthese* **121,** 29–54 (1999).

17. Husmeier, D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* **19,** 2271–2282 (2003).

18. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7,** 601–620 (2000).

19. Sachs, K., Perez, O. & Pe'er, D. Causal protein-signaling networks derived from multiparameter single-cell data. *Science.* **308,** 523–529 (2005).

20. Spirtes, P., Glymour, C.N. & Scheines, R. *Causation, Prediction, and Search* 2nd edn. (MIT Press, 2000).

21. Cantone, I. *et al.* A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell* **137,** 172–81 (2009).

22. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9,** 796–804 (2012).

23. Stolovitzky, G., Monroe, D. & Califano, A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann. N. Y. Acad. Sci.* **1115,** 1–22 (2007).

24. Stolovitzky, G., Prill, R.J. & Califano, A. Lessons from the DREAM2 Challenges. *Ann. N. Y. Acad. Sci.* **1158,** 159–95 (2009).

25. Prill, R.J. *et al.* Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One* **5,** e9202 (2010).

26. Prill, R.J., Saez-Rodriguez, J., Alexopoulos, L.G., Sorger, P.K. & Stolovitzky, G. Crowdsourcing network inference: the DREAM predictive signaling network challenge. *Sci. Signal.* **4,** mr7 (2011).

27. Meyer, P. *et al.* Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. *BMC Syst. Biol.* **8,** 13 (2014).

28. Tibes, R. *et al.* Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol. Cancer Ther.* **5,** 2512–21 (2006).

29. Mertins, P. *et al.* Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Mol. Cell. proteomics* **13,** 1690–1704 (2014).

30. Derry, J.M.J. *et al.* Developing predictive molecular maps of human disease through community-based modeling. *Nat. Genet.* **44,** 127–130 (2012).

31. Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. *Proc. 23rd Int. Conf. Mach. Learn. - ICML '06* 233–240 (2006).

32. Costello, J.C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32,** 20–23 (2014).

33. Margolin, A.A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5,** 181re1 (2013).

34. Cerami, E.G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39,** D685–690 (2011).

35. Wang, H. & Song, M. Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. *R J.* **3,** 29–33 (2011).

36. Hill, S.M., Nesser, N.K. *et al.* Context-specificity in causal signaling networks revealed by phosphoprotein profiling. (Submitted).

37. Maathuis, M.H., Colombo, D., Kalisch, M. & Bühlmann, P. Predicting causal effects in large-scale systems from observational data. *Nat. Methods* **7,** 247–8 (2010).

38. Olsen, C. *et al.* Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics* **103,** 329–336 (2014).

39. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–504 (2003).

Figure Legends

Figure 1

Causal networks. The network inference sub-challenge focused on causal relationships between nodes. (**a**) A directed edge (or link) in a causal network carries the interpretation that inhibition of the parent node (A) can change abundance of the child node (B) (the change could be up or down, here the latter is shown). (**b**) Causal edges, as used here, may represent direct effects or indirect effects that occur via unmeasured intermediate nodes. If node A causally influences node B via a measured node C, the causal network should contain edges from A to C and C to B, but no edge from A to B (top). However, if node C were not measured (and not part of the network), the causal network should contain an edge from A to B (bottom). Note that in both these cases inhibition of node A would lead to a change in node B. (**c**) Causal edges may depend on biological context. In the example shown, there is a causal edge from A to B in Context 1, but not in Context 2 (line colors are as in **a**). (**d**) Correlation and causation. In the example shown, nodes A and B are correlated due to regulation by the same node (C). However, in this example no sequence of mechanistic events links A to B, and thus inhibition of A does not change the abundance of B (line colors are as in **a**). Therefore, despite the correlation, there is no causal edge from A to B.

Figure 2

The HPN-DREAM network inference challenge: overview of experimental data tasks and causal assessment strategy. (**a**) Protein data were obtained from four cancer cell lines under eight ligand stimuli (data as described in Hill, Nesser *et al.*[36]). For each of the 32 resulting contexts, participants were provided training data comprising time-courses for ~45 phosphoproteins under three different kinase inhibitors and a control (DMSO). The sub-challenge 1 experimental data task (SC1A) asked participants to infer signaling networks specific to each context. In SC2A the aim was to predict context-specific molecular time-courses. In both cases, submissions were assessed using held-out, context-specific test data that were obtained under an unseen intervention (inhibition of the kinase mTOR). Each sub-challenge also comprised a companion *in silico* data task (SC1B/SC2B; text, Online Methods and **Supplementary Fig. 1**). (**b**) Networks submitted to the experimental data task (SC1A) were assessed causally in terms of agreement with the interventional test data. For each context, the set of nodes that changed under mTOR inhibition was identified ("gold-standard" causal descendants of mTOR; see text and Online Methods). In the example shown, node X is a descendant of mTOR while Y is not. (**c**) In submitted, context-specific networks, predicted descendants of mTOR were identified and compared with their experimentally-determined "gold-standard" counterparts. This gave true and false positive counts and a (context-specific) AUROC (area under the receiver operating characteristic curve). (**d**) In each context, teams were ranked by AUROC score and mean rank across contexts gave the final rankings (**Table 1**).

Figure 3

Network inference sub-challenge (SC1) results. (**a**) Heatmap showing assessment scores (AUROC) in each of the 32 *(cell line, stimulus)* contexts for the 74 teams that made submissions to the experimental data task (teams ordered by final ranking). (**b**) Scatter plot comparing scores in experimental and *in silico* data tasks. Each square represents a team, with color indicating whether prior information was used for the experimental data task and red border indicating that a different method was used in each task. Numerical annotation indicates ranks for the top ten teams under a combined score (see text; three teams were jointly ranked third under this score). (**c**,**d**) Results of crowdsourcing for the experimental data task. Aggregate submission networks were formed by combining, for each context, networks from top scoring (**c**) or randomly selected (**d**) teams (Online Methods). Blue line shows performance of the aggregate submission versus number of teams aggregated, with

red circles depicting individual team scores. Dashed black line indicates the result of aggregating all submissions. Results in **d** are mean values over 100 iterations of random selection (error bars indicate s.d.). See **Supplementary Figure 8** for analysis of the *in silico* data task. (**e**,**f**) Performance by method type for the experimental (**e**) and *in silico* (**f**) data tasks. Final rank appears above each bar and the gray line shows mean performance of random predictions. Note that some teams used a different approach for each task or only participated in a single task (**Supplementary Table 2**).

Figure 4

Network inference sub-challenge experimental data task (SC1A): role of prior information. (**a**) Tukey-style box plots over mean AUROC scores for teams that did/did not use a prior network. P-value calculated by a Wilcoxon rank-sum test (n = 18). (**b**) An aggregate prior network (text and Online Methods) was combined with networks inferred by two top-performing methods: (i) the "PropheticGranger" approach (top performer in SC1A when combined with a network prior) and (ii) the "FunChisq" approach (top performer in SC1B). Horizontal axis indicates relative contribution of the aggregate prior (zero indicates no contribution and one indicates aggregate prior alone) and the vertical axis is the score of the resulting networks. The blue line indicates performance when combining the aggregate prior with randomly generated networks (mean performance over 30 random networks shown, with shaded region indicating standard deviation). The dashed black line shows the mean AUROC score achieved by the top-performing team in the experimental data task. Error bars indicate s.e.m. (**c**) Top: Tukey-style box plots over AUROC scores for the top 25 performers for each context. The green triangles and purple circles indicate performance of the aggregate submission and the aggregate prior respectively. Bottom: Receiver operating characteristic (ROC) curves for two contexts where a difference in performace was observed between the aggregate submission and aggregate prior.

Figure 5

Aggregate submission networks for the network inference experimental data task (SC1A). (**a**) Aggregating all submissions gave a network for each of 32 *(cell line, stimulus)* contexts; the network for cell line MCF7 under HGF stimulus is shown. Line thickness corresponds to edge weights (number of edges shown set to equal number of nodes). Black solid (red dashed) lines indicate edges that are present (not present) in the aggregate prior network (obtained by placing a threshold of 0.1 on edge weights). Green/blue nodes are descendants of mTOR in the network shown; green nodes are true positives with respect to the test data while blue nodes are false positives (**Figs. 2b,c** and **Supplementary Fig. 2**). Network generated using Cytoscape[39]. (**b**) Principal components analysis (PCA) was applied to the context-specific aggregate submission networks (Online Methods) with the 32 contexts colored by cell line (PCA was applied to vectors comprising edge scores for the 32 aggregate networks).

Table 1

Summary of highly-ranked submissions to the network inference sub-challenge (SC1). 41 of the 80 teams that made submissions to the sub-challenge provided information regarding their methodologies. Methods were categorized based on this information (see also **Fig. 3d** and **Supplementary Note 5**). SC1A denotes experimental data task and SC1B denotes *in silico* data task. Only teams ranked within the top 5 in either SC1A, SC1B or under a combined score are shown. Teams ordered by SC1A rank. The full table with all teams is provided in **Supplementary Table 2**, together with additional metrics and characterizations for each method. Detailed method descriptions can be found in **Supplementary Note 1**.

Anonymized Name	Summary[1]	Classification	SC1A mean AUROC	SC1B AUROC	Final rank			Used prior network?[2]
					SC1A	SC1B	SC1A/B combined	
Team1 (top-performer)	"PropheticGranger with heat diffusion prior": an extension of L1-penalized Granger causality constructed specifically to consider "future data", combined with a prior derived from known biological pathways.	Linear regression	0.78	0.56	**1**	28	11	Yes (Team2)
Team2	For experimental data, a biological prior is created by applying a simulated heat diffusion process to the constituent pathways from Pathway Commons. For *in silico* data, the network inference method ARACNE is used.	Prior only / Pairwise score	0.77	0.46	2	65	24	Yes
Team3	Edges are removed from a literature-based network by considering time-lagged correlation between phosphoprotein pairs, fold-changes in abundance through time and results from the time-course prediction challenge.	Pairwise score	0.76	0.56	3	23	9	Yes
Team4	An ensemble approach combining prior knowledge with results from L1-penalized Granger causality and the GENIE3 algorithm.	Ensemble	0.75	0.62	4	11	3	Yes (Team2)
Team5	A random forest classifier, with a literature-derived network used as a "gold standard", predicts existence of edges from several measures of pair-wise association, calculated using multiple sources of data	Nonlinear regression	0.73	0.60	5	13	6	Yes
Team6	Network topology and parameters are optimized by minimizing an objective function based on a linear Ordinary Differential Equation (ODE) model, using a greedy search starting from a knowledge-based network topology.	ODEs	0.72	0.66	6	7	2	Yes
Team7 (top-performer)	"FunChisq": a functional chi-square test to infer causal network topology based on nonparametric functional dependency from data discretized by optimal k-means clustering for each variable.	Pairwise score[3]	0.71	0.76	7	**1**	**1**	No
Team10	A regression approach using a stationary Markov assumption and truncated singular value decomposition to predict network structure.	Linear regression	0.68	0.66	10	5	3	No
Team11	Boruta, a wrapper feature selection method, utilizing random ferns classifier as an importance source, is used to perform feature selection for each node in the network.	Nonlinear regression	0.64	0.73	13	2	3	No
Team17	The GENIE3 algorithm was used to infer weights and directions for regulatory edges between phosphoproteins, with incorporation of a prior knowledge network.	Nonlinear regression	0.61	0.67	26	4	12	Yes (Team2)
Team18	Network inference is formulated as a linear programming problem designed specifically for perturbation time series data and models signaling as information flow.	Other	0.57	0.68	29	3	14	No

[1]References to prior networks in the "Summary" column apply to the experimental data task only.
[2]SC1A only. Several teams used the prior network submitted by Team2; these teams are indicated by the text "(Team2)".
[3]Team7 chose to consider only pairwise interactions in their application of FunChisq to the challenge data, but the approach is also able to consider many-to-one interactions and this is implemented in the code provided.

The HPN-DREAM Consortium

Bahman Afsari[11], Rami Al-Ouran[26], Bernat Anton[27], Tomasz Arodz[28], Omid AskariSichani[29], Neda Bagheri[30], Noah Berlow[31], Alexander J. Bisberg[16], Adrian Bivol[8], Anwesha Bohler[32], Jaume Bonet[27], Richard Bonneau[33,34,35], Gungor Budak[32], Razvan Bunescu[26], Mehmet Caglar[36], Binghuang Cai[37], Chunhui Cai[37], Daniel E. Carlin[8], Azzurra Carlon[38], Lujia Chen[37], Mark F. Ciaccio[30], Thomas Cokelaer[5], Gregory Cooper[37], Chad J. Creighton[39], Seyed-Mohammad-Hadi Daneshmand[29], Alberto de la Fuente[40], Barbara Di Camillo[38], Ludmila V. Danilova[11,12], Joyeeta Dutta-Moscato[37], Kevin Emmett[41], Chris Evelo[32], Mohammad-Kasim H. Fassia[42], Alexander V. Favorov[11,12,13], Elana J. Fertig[11], Justin D. Finkle[43], Francesca Finotello[38], Stephen Friend[18], Xi Gao[28], Jean Gao[44], Javier Garcia-Garcia[27], Samik Ghosh[45], Alberto Giaretta[38], Kiley Graim[8], Joe W. Gray[2,3,4], Ruth Großeholz[46], Yuanfang Guan[10,19,20], Justin Guinney[18], Christoph Hafemeister[33], Oliver Hahn[46], Saad Haider[31], Takeshi Hase[45], Laura M. Heiser[2,3,4], Steven M. Hill[1], Jay Hodgson[18], Bruce Hoff[18], Chih Hao Hsu[47], Chenyue W. Hu[16], Ying Hu[47], Xun Huang[48], Mahdi Jalili[29], Xia Jiang[37], Tim Kacprowski[49], Lars Kaderali[50,51], Mingon Kang[44], Venkateshan Kannan[52], Michael Kellen[18], Kaito Kikuchi[45], Dong-Chul Kim[53], Hiroaki Kitano[45], Bettina Knapp[50,54], George Komatsoulis[47], Heinz Koeppl[6,25], Andreas Krämer[55], Miron Bartosz Kursa[56], Martina Kutmon[32], Wai Shing Lee[11], Yichao Li[26], Xiaoyu Liang[26], Zhaoqi Liu[57], Yu Liu[58], Byron L. Long[16], Songjian Lu[37], Xinghua Lu[37], Marco Manfrini[38], Marta R. A. Matos[50], Daoud Meerzaman[47], Gordon B. Mills[17], Wenwen Min[57], Sach Mukherjee[1,23,24,25], Christian Lorenz Müller[33,34], Richard E. Neapolitan[59], Nicole K. Nesser[7], David P. Noren[16], Thea Norman[18], Baldo Oliva[27], Stephen Obol Opiyo[60], Ranadip Pal[31], Aljoscha Palinkas[61], Evan O. Paull[8], Joan Planas-Iglesias[27], Daniel Poglayen[27], Amina A. Qutub[16], Julio Saez-Rodriguez[5,22], Francesco Sambo[38], Tiziana Sanavia[38], Ali Sharifi-Zarchi[62], Janusz Slawek[28], Artem Sokolov[8], Mingzhou Song[9], Paul T. Spellman[7], Adam Streck[61], Gustavo Stolovitzky[21], Sonja Strunz[40], Joshua M. Stuart[8], Dane Taylor[14,15], Jesper Tegnér[52], Kirste Thobe[61], Gianna Maria Toffolo[38], Emanuele Trifoglio[38], Michael Unger[6], Qian Wan[31], Haizhou Wang[9,25], Lonnie Welch[26], Chris K. Wong[8], Jia J. Wu[43], Albert Y. Xue[30], Ryota Yamanaka[45], Chunhua Yan[47], Sakellarios Zairis[63], Michael Zengerling[46], Hector Zenil[52], Shihua Zhang[57], Yang Zhang[9,25], Fan Zhu[10], Zhike Zi[48]

26. Ohio University, School of Electrical Engineering and Computer Science, Russ College of Engineering and Technology, Athens, OH, USA
27. Universitat Pompeu Fabra, Structural Bioinformatics Group (GRIB/IMIM), Departament de Clències Experimentals i de la Salut, Barcelona, Catalonia, Spain
28. Virginia Commonwealth University, School of Engineering, Department of Computer Science, Richmond, VA, USA
29. Sharif University of Technology, Computer Engineering Department, Tehran, Iran
30. Northwestern University, Chemical & Biological Engineering, Evanston, IL, USA
31. Texas Tech University, Department of Electrical and Computer Engineering, Lubbock, TX, USA
32. Maastricht University, Department of Bioinformatics - BiGCaT, Maastricht, The Netherlands
33. New York University, Department of Biology, Center for Genomics & Systems Biology, New York, NY, USA
34. New York University, Courant Institute of Mathematical Sciences, New York, NY, USA
35. Simons Foundation, Simons Center for Data Analysis, New York, NY, USA
36. Texas Tech University, Department of Physics, Lubbock, TX, USA
37. University of Pittsburgh, Biomedical Informatics, Pittsburgh, PA, USA
38. University of Padova, Department of Information Engineering, Padova, Ital
39. Baylor College of Medicine, Department of Medicine and Dan L. Duncan Cancer Center Division of Biostatistics, Houston, TX, USA
40. Leibniz-Institute for Farm Animal Biology, Department of Biomathematics and Bioinformatics, Institute of Genetics and Biometry, Dummerstorf, Germany

41. Columbia University, Department of Physics, New York, NY, USA
42. Northwestern University, Biomedical Engineering, Evanston, IL, USA
43. Northwestern University, Interdepartmental Biological Sciences, Evanston, IL, USA
44. University of Texas at Arlington, Computer Science and Engineering, Arlington, TX, USA
45. The Systems Biology Institute, Research Department, Tokyo, Japan
46. University Heidelberg COS Heidelberg, BioQuant (BQ0018), Department of Modeling Biological Processes, Heidelberg, Germany
47. National Cancer Institute, Center for Biomedical Informatics & Information Technology, Bethesda, MD, USA
48. University of Freiburg, BIOSS Centre for Biological Signalling Studies, Freiburg, Germany
49. Interfaculty Institute for Genetics and Functional Genomics, University Medicine and Ernst-Moritz-Arndt University Greifswald, Department of Functional Genomics, Greifswald, Germany
50. Technische Universität Dresden, Institute for Medical Informatics and Biometry, Medical Faculty Carl Gustav Carus, Dresden, Germany
51. University Medicine Greifswald, Institute for Bioinformatics, Walther-Rathenau-Str. 48, 17475 Greifswald, Germany
52. Karolinska Institute, Unit of Computational Medicine, Science for Life Laboratory (SciLifeLab), Center for Molecular Medicine, Department of Medicine, Solna, Stockholm, Sweden
53. University of Texas- Pan American, Department of Computer Science, TX, USA
54. Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany
55. QIAGEN, Redwood City, CA, USA
56. University of Warsaw, Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw, Poland
57. National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
58. St. Jude Children's Research Hospital, Department of Computational Biology, Memphis, TN, USA
59. Northwestern University Feinberg School of Medicine, Division of Biomedical Informatics, Department of Preventive Medicine, Chicago, IL, USA
60. Ohio State University, Molecular and Cellular Imaging Center-Columbus, Columbus, OH, USA
61. Freie Universitât Berlin, Department Of Mathematics and Computer Science, Berlin, Germany
62. Royan Institute for Stem Cell Biology and Technology, ACECR, Department of Stem Cells and Developmental Biology at Cell Science Research Center, Tehran, Iran
63. Columbia University, Center for Computational Biology and Bioinformatics, New York, NY, USA

Online Methods

Challenge data

The HPN-DREAM network inference challenge comprised three sub-challenges: causal network inference (SC1), time-course prediction (SC2) and visualization (SC3). Each of SC1 and SC2 consisted of two tasks, one based on experimental data (SC1A/SC2A) and the other based on *in silico* data (SC1B/SC2B).

Experimental data. The experimental data and associated components of the challenge are outlined in **Figure 2a**. Provided for the challenge were protein data from four breast cancer cell lines (UACC812, BT549, MCF7, BT20), All cell lines were acquired from ATCC, authenticated by STR analysis, and tested for mycoplasma contamination. The cell lines

were chosen because they represent the major subtypes of breast cancer (basal, luminal, claudin-low, and HER2-amplified) and are known to have different genomic aberrations[40–42]. Each cell line sample was stimulated with 8 ligands (serum, PBS, EGF, Insulin, FGF1, HGF, NRG1, IGF1). We refer to each of the 32 possible combinations of cell line and stimulus as a biological context. For each context, data comprised time-courses for total proteins and post-translationally modified proteins, obtained under four different kinase inhibitors and a DMSO control. Full details of sample preparation, data generation, quality control, and pre-processing steps can be found in Hill, Nesser *et al*.[36] and on the Synapse[30] webpages describing the challenge (https://www.synapse.org/HPN_DREAM_Network_Challenge). In brief, cell lines were serum-starved for 24 hours and then treated for two hours with an inhibitor (or combination of inhibitors or DMSO vehicle alone). Cells were then either harvested (0 time point) or stimulated by one of the eight stimuli for 5, 15, 30, or 60 minutes, or 2, 4, 12, 24, 48, or 72 hours prior to protein harvest and analysis by reverse-phase protein array (RPPA) at MD Anderson Cancer Center Functional Proteomics Core Facility (Houston, Texas).

RPPA is an antibody-based assay that provides quantitative measurements of protein abundance[28,43]. The MD Anderson RPPA core facility maintains and updates a standard antibody list on the basis of antibody quality control as well as a variety of other factors, including scientific interest. Antibodies available for use in this assay are therefore enriched for components of receptor tyrosine kinase (RTK) signaling networks and cancer-related proteins. For each cell line, we used the standard antibody list available at the time the assays were performed. 183 antibodies were used to target total (n = 132), cleaved (n = 3) and phosphoproteins (n = 48; **Supplementary Table 1**). As part of the RPPA pipeline, quality control was performed to identify slides with poor antibody staining. Antibodies with poor quality control scores were excluded from the dataset. During the challenge period, it became known to challenge organizers that several further antibodies were of poor quality. Participants were advised not to include the associated data in their analyses and these data were excluded from the scoring process. Measurements for each sample were corrected for protein loading and several outlier samples with large correction factors were identified and removed. The UACC812 data were split across two batches. A batch-normalization procedure was applied[36] to enable the data from the two batches to be combined. The experimental data used in the challenge is a subset of the data reported in Hill, Nesser *et al*.[36].

The inhibitors were chosen because they target key components of the RTK signaling cascades assessed by the RPPA assay and are also relevant to breast cancer. Participants were provided with a training dataset consisting of data for four out of the five inhibitor regimes (DMSO, PD173074 (FGFRi), GSK690693 (AKTi), GSK690693+GSK1120212 (AKTi+MEKi)). Note that there was no training data available for the AKTi+MEKi inhibitor regime for cell lines BT549 (all stimuli) and BT20 (PBS and NRG1 stimuli). Data for the remaining inhibitor (AZD8055 (mTORi)) formed a test dataset, unseen by participants and used to evaluate submissions to the challenge.

The focus of the challenge was on short-term phosphoprotein signaling events and not on medium-to-long-term changes over hours and days (e.g. re-wiring of networks due to epigenetic changes arising from prolonged exposure to an inhibitor). Therefore the training data consisted only of phosphoprotein data (~45 phosphoproteins for each cell line) up to and including the four-hour time point; in the challenge this dataset was referred to as the "Main" dataset. In case some participants found the additional data useful, measurements for the remaining antibodies and time points were also made available in a "Full" dataset. The test data (and challenge scoring) also focused only on phosphoproteins up to and including the four-hour time point. At the time of the challenge all data were unpublished (the training dataset was made available to participants through the Synapse platform).

In silico data. The *in silico* data and associated components of the challenge are outlined in **Supplementary Figure 1**. Simulated data were generated from a nonlinear ordinary differential equation (ODE) model of the ERBB signaling pathway. Specifically, the model was an extended version of the mass action kinetics model developed by Chen *et al.*[12]. Training data were simulated for 20 network nodes (**Supplementary Fig. 4**; 14 phosphoproteins, two phospholipids, GTP-bound RAS and three dummy nodes that were unconnected in the network) under two ligand stimuli (each at two concentrations; applied individually and in combinations) and under three inhibitors targeting specific nodes in the network or no inhibitor. Mirroring the experimental data, inhibitors were applied prior to ligand stimulation at t=0. Time-courses consisted of 11 time points (0, 1, 2, 4, 6, 10, 15, 30, 45, 60, 120 minutes) and three technical replicates were provided for each sample. A measurement error model was developed to reflect the antibody-based readout of RPPA assays and its technical variability. Node names were anonymized to prevent prior information being used to trivially reconstruct the network. Further details of the simulation model can be found in **Supplementary Note 8**.

An *in silico* test dataset was also generated to assess submissions to the time-course prediction sub-challenge and consisted of time-courses for each node and stimulus, under *in silico* inhibition of each network node in turn. After the final team rankings for the *in silico* data task were calculated, two minor issues concerning the *in silico* test data were discovered. The issues were corrected, test data were regenerated and final rankings and final leaderboards were updated. The top-performing teams remained unchanged by this update. See **Supplementary Note 8** for further details.

Challenge questions and design

For the network inference sub-challenge experimental data task, participants were asked to use the training data to learn 32 signaling networks, one for each of the *(cell line, stimulus)* contexts. Networks had to contain nodes for each phosphoprotein in the training data (the node set therefore varied depending on cell line) and network edges had to be directed (but unsigned). The networks were expected to describe causal edges and this was reflected in the scoring (see below). A causal edge was defined as one where inhibition of the parent node can result in a change in the abundance of the child node that is not fully mediated via any other measured node (but the influence can take place via unmeasured nodes; **Fig. 1**). Participants were asked to submit confidence scores (between 0 and 1) for each possible directed edge in each network. Node names were *not* anonymized for the experimental data task and participants were allowed to use pre-existing biological information (e.g. from literature and online databases) in their analyses.

For the network inference sub-challenge *in silico* data task, participants were asked to infer a single network with 20 nodes (one for each variable in the training data) and directed edges corresponding to predicted causal relationships between the nodes. Submissions comprised a set of confidence scores for each possible directed edge in the network.

For the time-course prediction sub-challenge, participants were tasked with predicting time-courses under interventions not contained in the training dataset. For the experimental data task, predictions were requested for five test kinase inhibitors (participants were informed of the inhibitor targets). For each inhibitor, time-courses consisting of seven time points (as in the training data) had to be predicted for each of the 32 contexts and for all phosphoproteins (except those targeted by the inhibitor). The *in silico* data task proceeded in an analogous fashion, with participants asked to predict time-courses under inhibition of each of the 20 nodes in turn. Predicted time-courses were required for each node for each of the eight stimulus contexts.

The visualization sub-challenge asked participants to devise novel approaches to represent the dataset provided with the challenge. The submission format was a schematic mock-up of the visualization.

The challenge was run over a period of three months. For the network inference and time-course prediction sub-challenges participants were able to make submissions and obtain feedback via a leaderboard on a weekly basis (**Supplementary Note 9**). The frequency of feedback was chosen so as to obtain a balance between actively engaging participants and avoiding overfitting of models to the test data. To address this overfitting issue, other DREAM challenges[33,44] used a second held-out test dataset for final scoring of submissions. However, this was not possible here due to the small number of inhibitor conditions in the data.

To incentivize participation, top-performing teams were awarded a modest cash prize (provided by the Heritage Provider Network), invitations to present results at a conference and co-author the paper describing the challenge, and (for SC1A only) the opportunity to have their method developed as a Cytoscape Cyni App[39,45]. See the Synapse pages describing the challenge (www.synapse.org/#!Synapse:syn1720047) and **Supplementary Note 7** for further details.

Scoring procedure for the network inference sub-challenge experimental data task
Interventional test data. For the experimental data task, we developed a scoring procedure that used held-out interventional data to assess networks submitted by participants. The procedure assessed the extent to which causal relationships encoded in network submissions agreed with causal information contained in the test data. Using the held-out mTOR inhibitor data, we identified those phosphoproteins that showed a salient change in abundance under the inhibitor, relative to DMSO control (**Fig. 2b**). Specifically, let $\mu_{i,c}^{D}$ and $\mu_{i,c}^{I}$ denote the mean abundance levels of phosphoprotein *i* for *(cell line, stimulus)* context *c* under DMSO control and mTOR inhibition respectively (mean values calculated over 7 time points on log-transformed data; any replicates at each time point were averaged prior to taking the mean). A paired *t*-test was used to assess whether $\mu_{i,c}^{D}$ is significantly different to $\mu_{i,c}^{I}$, resulting in a *p*-value $p_{i,c}$ for each phosphoprotein and context.

Some phosphoproteins show a clear stimulus response under DMSO, characterized by a marked increase and subsequent decrease in abundance over time (a "peak" shape). In these cases, a change in abundance due to the mTOR inhibitor may be observable only at intermediate time points. Since the paired *t*-test described above considers all time points, the effect may be masked. Therefore we used a heuristic to detect phosphoproteins with a peak shaped time-course under DMSO and re-performed the paired *t*-test over the intermediate time points within the peak only. The resulting *p*-value was retained if smaller than the original. For each context, a test is performed for each phosphoprotein. We corrected for multiple testing within each context using the median adaptive linear step-up procedure[46], resulting in *q*-values (FDR-adjusted *p*-values) $q_{i,c}$. Note that due to the heuristic step, the *q*-values $q_{i,c}$ should not be interpreted formally.

For each context, a phosphoprotein is determined to show a change under the mTOR inhibitor if the following two conditions are satisfied: (1) $q_{i,c} < 0.05$ and (2) $\left|\mu_{i,c}^{D} - \mu_{i,c}^{I}\right| > \sigma_{i,c}$ where $\sigma_{i,c}$ is the pooled replicate standard deviation for the DMSO and mTOR inhibitor data. The second condition acts as a conservative filter to ensure effect sizes are not small relative to replicate variation. We work under the assumption that mTOR inhibition will lead to changes in abundance of all descendants of mTOR in the underlying context-specific causal network (i.e. changes are observed in any node for which a directed path exists from mTOR to that node; this can include downstream nodes as well as those that are influenced via feedback loops within the timescale of the experiments). Then, the above procedure

results in context-specific "gold-standard" sets of causal descendants of mTOR $D_c^{\mathrm{GS}} = \{i | q_{i,c} < 0.05 \text{ and } |\mu_{i,c}^D - \mu_{i,c}^I| > \sigma_{i,c}\}$ (**Supplementary Fig. 2**).

The scoring metric. For each context *c*, we compared the "gold-standard" descendant set D_c^{GS} (obtained from the held-out test data) with predicted descendant sets obtained from context-specific networks submitted by participants (**Fig. 2c**). For context *c*, a submitted network consisted of edge confidence scores for each possible directed edge. Placing a threshold τ on edge scores resulted in a network structure consisting only of those edges with a score greater than τ and from this network we obtained a predicted set of descendants of mTOR (at threshold τ), denoted by $D_c^{\mathrm{pred}}(\tau)$. Comparing $D_c^{\mathrm{pred}}(\tau)$ with D_c^{GS} gave the number of predicted descendants that are correct (true positives; $TP(\tau)$) and the number of predicted descendants that are incorrect (false positives; $FP(\tau)$). Varying the threshold τ and plotting $TP(\tau)$ against $FP(\tau)$ resulted in a receiver operating characteristic (ROC) curve and the scoring metric was the area under this curve (normalized to be between zero and one; AUROC). For each team AUROC scores were calculated for each of the 32 contexts.

The statistical significance of AUROC scores was determined using simulated null distributions, generated by calculating AUROC scores for 100,000 random networks, each consisting of random edge scores (drawn independently from the uniform distribution on the unit interval [0,1]). Gaussian fits to the null distributions were used to calculate *p*-values. For each context, the set of *p*-values (across all teams) underwent multiple testing correction using the Benjamini-Hochberg FDR procedure. There were two contexts (BT549, NRG1 and BT20, insulin) for which no team achieved a statistically significant (FDR < 0.05) AUROC score (**Supplementary Fig. 7b**). These two contexts were therefore regarded as too challenging and were disregarded in the scoring procedure.

Teams were ranked within each context according to AUROC score. The resulting 30 rank scores for each team were then averaged to obtain a mean rank score. Final team rankings were obtained using mean rank scores (**Fig. 2d**).

During the challenge period, participants were informed only that submitted networks would be scored using test data obtained under interventions not present in the training data, but details of the scoring procedure and the identity, nature and number of interventions in the test data were not revealed. Note that participants knew identities of inhibitors in the training data.

Gold-standard network and scoring metric for the network inference sub-challenge *in silico* data task

The true causal network underlying the variables in the *in silico* data was obtained from the data-generating nonlinear ODE model (**Supplementary Fig. 4**). However, deriving the causal network from the equations was non-trivial due to the model containing more variables than the 20 variables present in the challenge data and variables appearing in the model in complexes. Details of how the causal network was obtained can be found in **Supplementary Note 8**.

Each team submitted a single network consisting of a set of edge scores. This was compared directly to the gold-standard causal network to produce an ROC curve (by calculating the number of true positive and false positive edges at various edge score thresholds) and AUROC was used as the scoring metric. Self-edges were not considered for scoring. Statistical significance of AUROC scores was determined analogously to the experimental data task.

Alternative scoring metrics for the network inference sub-challenge

We used AUROC as the scoring metric for the network inference sub-challenge but we note that alternative metrics could have been used. In particular, the area under the precision-recall curve (AUPR) is often used when there is an imbalance between the number of positives and negatives in the gold standard[31]. While many contexts in the experimental data task had a reasonable balance (median ratio of negatives to positives of 1.71), some contexts had many more negatives than positives and there was also an imbalance for the *in silico* data task (ratio of negatives to positives of 4.14; **Supplementary Fig. 5**). Therefore AUPR could have been an appropriate choice in several cases. For this reason, at the end of the challenge period we performed comparisons of final team rankings (obtained using AUROC) against rankings obtained using AUPR, or a combination of both AUROC and AUPR (**Supplementary Fig. 6**). For the experimental data task, the AUROC-based rankings showed good agreement with those obtained under either alternative metric. Agreement was not as strong for the *in silico* data task, but still reasonable with all metrics resulting in the same top performer. Furthermore, of the top ten teams under AUROC, only 2 are outside the top 10 under AUPR and they are ranked 12 and 13. Similarly, only 2 of the top 10 teams under AUPR are not in the top 10 under AUROC, and they are ranked 11 and 12. For openness and transparency, scores and rankings based on AUPR and the combination metric were included in the final leaderboards (available through Synapse: https://www.synapse.org/HPN_DREAM_Network_Challenge; combination metric scores are also included in **Supplementary Table 2**).

Scoring metric for the time-course prediction sub-challenge

For both experimental data and *in silico* data, predictions of context-specific time-courses under inhibitors not contained in the training data were directly compared against context-specific test data obtained under the corresponding inhibitor. Prediction accuracy was quantified using root mean squared error (RMSE) with comparisons made on log-transformed data after averaging of replicates. RMSE scores were calculated separately for parts of the data that could potentially be on different scales. We refer to each portion of the data where an RMSE score was calculated as a “data block”. Teams were ranked within each “data block” and a mean rank calculated to obtain a final ranking. Some blocks of data, where no team achieved a statistically significant score, were disregarded in the scoring procedure (**Supplementary Tables 5** and **6**; FDR < 0.05). Full details of scoring appear in **Supplementary Note 6**.

Visualization sub-challenge scoring

HPN-DREAM challenge participants scored submitted visualization proposals. 36 participants cast votes by assigning ranks (from 1 to 3) to their three favorite submissions (the remaining submissions were all assigned a rank of 4). Teams were then ranked according to mean rank across the 36 votes (**Supplementary Fig. 11**).

Robustness of ranking under subsampling

To ensure team rankings were robust in the network inference and time-course prediction sub-challenges, we performed a subsampling analysis in which, for each of 100 iterations, 50% of the test data were removed at random and rankings of submissions were recalculated using the remaining test data. Team *A* was considered to be robustly ranked above team *B* if the former outranked the latter in at least 75% of iterations.

For the network inference sub-challenge experimental data task, test data were subsampled by either (i) removing 50% of the phosphoproteins for each *(cell line, stimulus)* context when making comparisons between “gold-standard” and predicted descendant sets (**Supplementary Fig. 12a**), or (ii) by removing 50% of the contexts (i.e. scoring was based on 15 contexts instead of 30; **Supplementary Fig. 12b**). The top team (Team1) outranked the team ranked second (Team2) in 76% and 97% of iterations for subsampling methods (i) and (ii) respectively. For the network inference sub-challenge *in silico* data task, 50% of the edges (and non-edges) in the gold-standard network were used for scoring (**Supplementary**

Fig. 12c). The top scoring performer (Team7) had a higher AUROC score relative to the team ranked second (Team11) in 89% of the subsampling iterations.

For the experimental and *in silico* data tasks comprising the time-course prediction sub-challenge, test data were subsampled by either (i) removing 50% of the "data blocks" (see above), or (ii) by subsampling 50% of the data points within each "data block". For the experimental data task, the top-ranked team (Team44) outranked the team ranked second (Team42) in 90% and 54% of iterations for subsampling methods (i) and (ii) respectively. Due to the 75% threshold not being met for one of the subsampling methods, Team44 was not regarded as being ranked robustly above Team42. Team42 outranked the team ranked third (Team10) in 60% and 70% of iterations and so, again, the ranking was not regarded as robust. However, Team10 was robustly ranked above the team ranked fourth (93% and 94% of iterations). Team44 was not eligible to be named as a top-performer due to an incomplete submission (**Supplementary Note 7**) and so the teams ranked second and third (Team42 and Team10) were named as top-performers. For the *in silico* data task, the top team (Team34) outranked the team ranked second in 95% and 100% of iterations for subsampling methods (i) and (ii) respectively.

Crowdsourced analyses: aggregate submission networks and aggregate prior network

Aggregate submission networks were obtained by integrating predicted networks across all teams (to avoid bias, a filtering process was used to remove correlated submissions from the aggregation; 66 and 57 teams formed the aggregate networks for the experimental and *in silico* data tasks respectively; **Supplementary Note 10** and **Supplementary Table 2**). For the experimental data task, an aggregate network was formed for each of the 32 contexts. Each aggregate submission network consisted of a set of edge scores, calculated by taking the mean of scores submitted by teams for each edge. To ensure edge scores were comparable across teams, scores for each team were scaled prior to aggregation so that the maximum edge score (across all 32 contexts for the experimental data task) had a value of one.

For the experimental data task, an aggregate prior network was formed in an analogous manner to the aggregate submission networks, using 10 prior networks provided by teams (the prior network submitted by Team2 was also used by several other teams, but was only included once in the aggregation; **Supplementary Table 2**). Individual prior networks and therefore the aggregate prior network were not context-specific.

Principal component analysis (PCA) of context-specific aggregate submission networks

The 32 context-specific aggregate submission networks for the network inference sub-challenge experimental data task were combined into a matrix E of edge scores where columns correspond to contexts and rows correspond to edges (only network nodes common to all contexts were considered for this analysis). Each row of matrix E contains the scores for a specific edge in each of the contexts. PCA was performed on this matrix using the MATLAB function *princomp*.

Web-based community resource

A community resource has been made available through the Synapse platform at https://www.synapse.org/HPN_DREAM_Network_Challenge, under the section titled "HPN-DREAM Community Resource". This resource includes: all challenge data, participant submissions, participant code, participant prior networks and crowdsourced aggregate networks. Code for scoring submissions is available as part of the *DREAMTools* software package[47] (**Supplementary Note 11**).

References

40. Neve, R.M. *et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10,** 515–27 (2006).

41. Garnett, M.J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483,** 570–575 (2012).

42. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–307 (2012).

43. Hennessy, B.T. *et al.* A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Non-microdissected Human Breast Cancers. *Clin. Proteomics* **6,** 129–151 (2010).

44. Eduati, F. *et al.* Prediction of human population responses to toxic compounds by a collaborative competition. *Nat. Biotechnol.* **33,** 933–940 (2015).

45. Guitart-Pla, O., Kustagi, M., Rügheimer, F., Califano, A. & Schwikowski, B. The Cyni framework for network inference in Cytoscape. *Bioinformatics* **31**, 1499–1501 (2015).

46. Benjamini, Y., Krieger, A.M. & Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93,** 491–507 (2006).

47. Cokelaer, T. *et al.* DREAMTools: a Python package for scoring collaborative challenges [version 1; referees: 3 approved with reservations]. *F1000Research* **4,** 1030 (2015).

Competing financial interests

The authors declare no competing financial interests.

Figure-1 (Mukherjee)

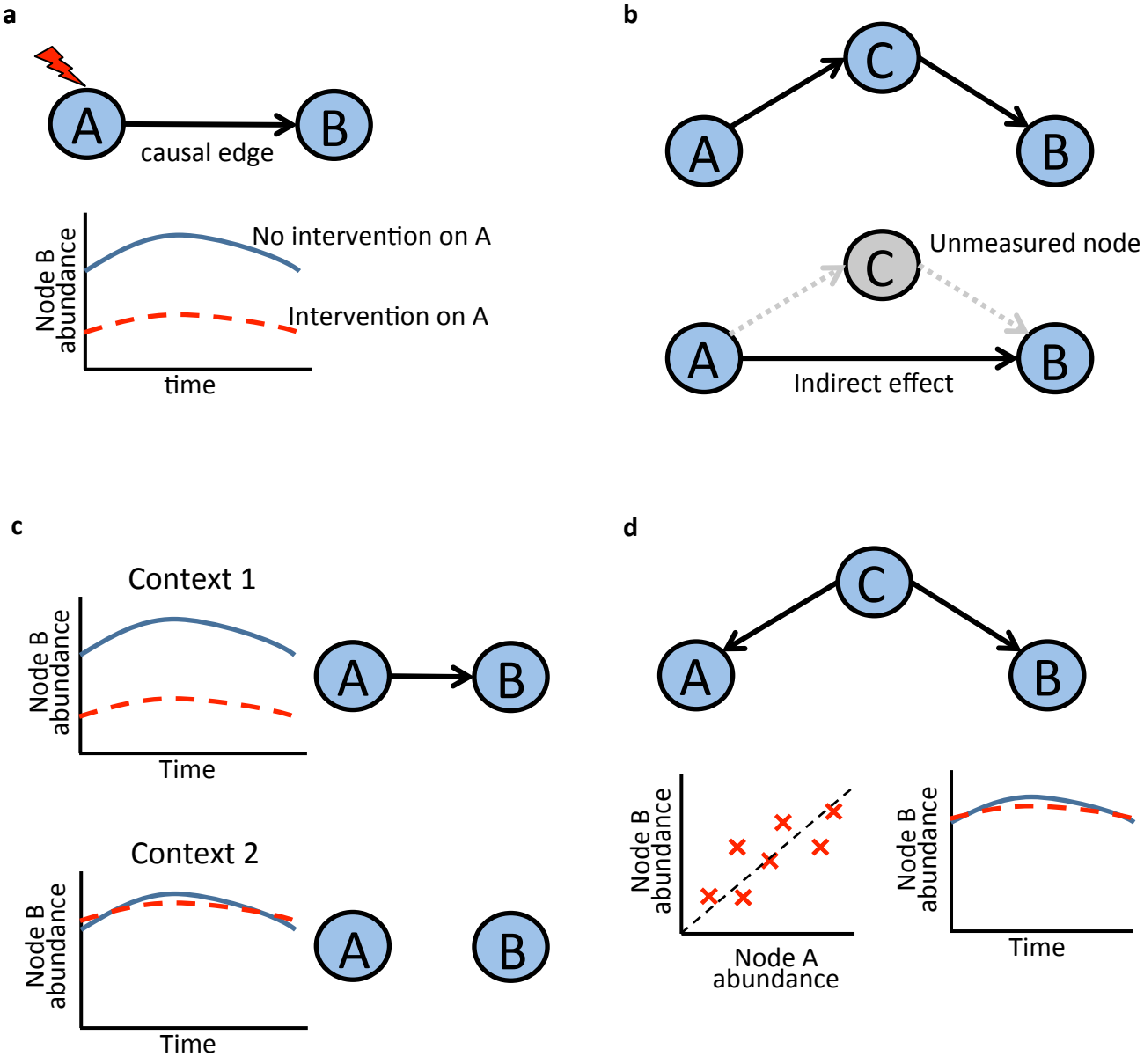
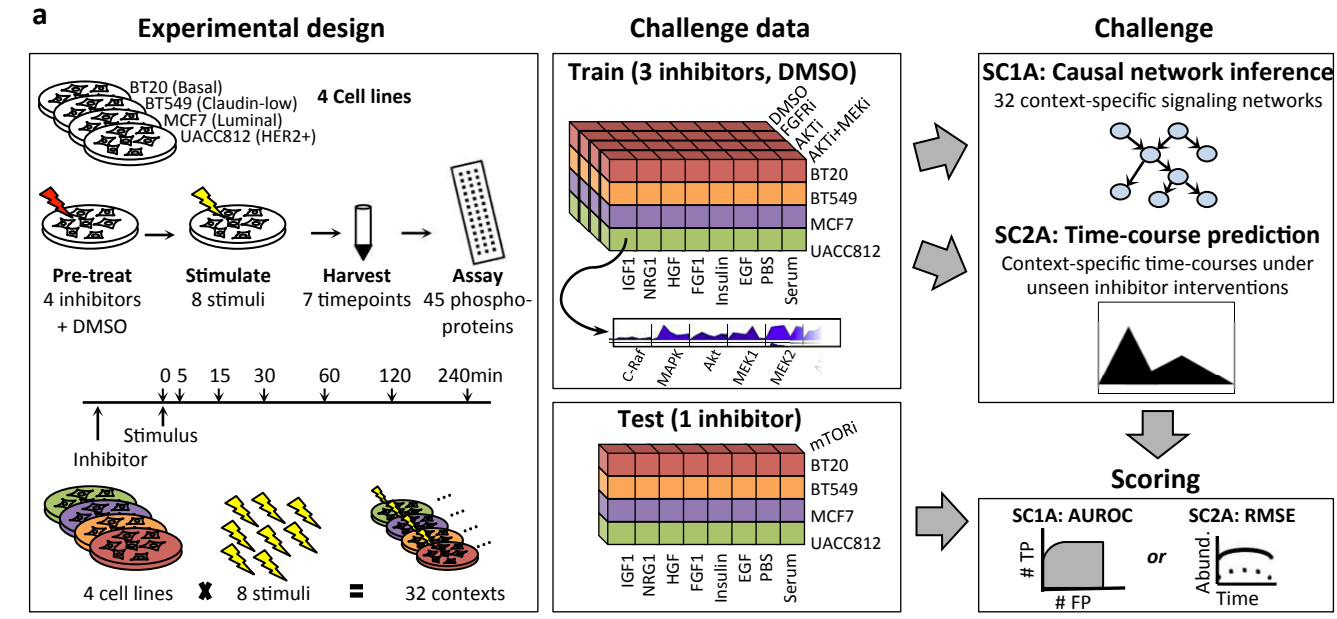
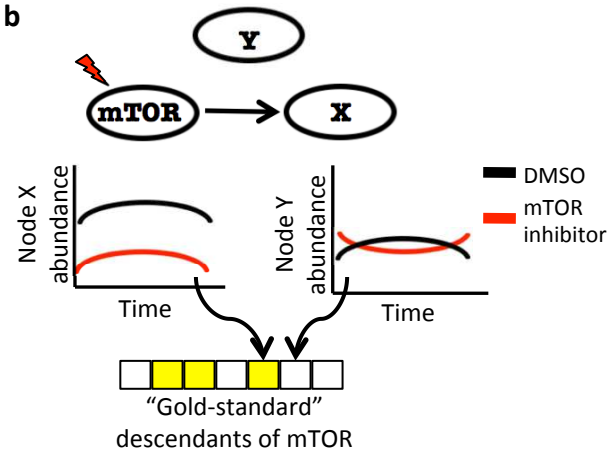
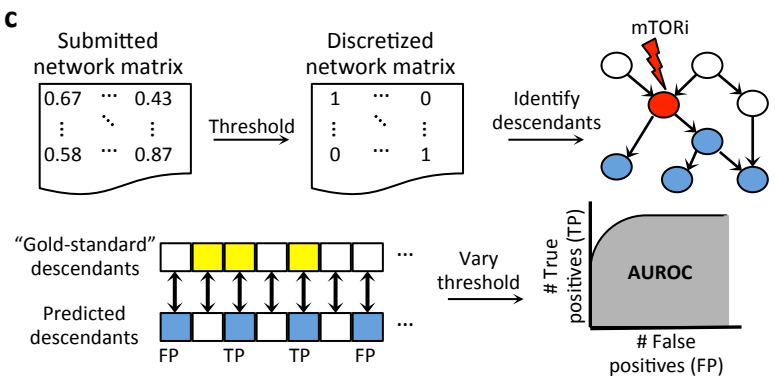
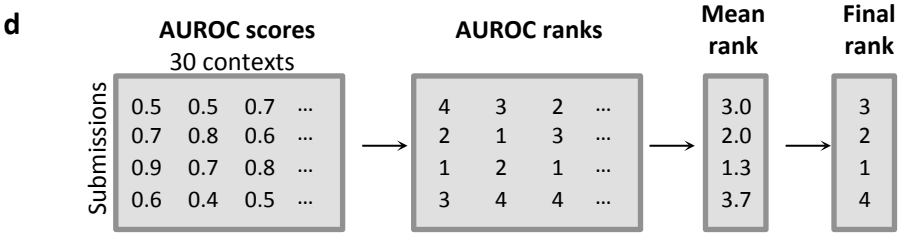


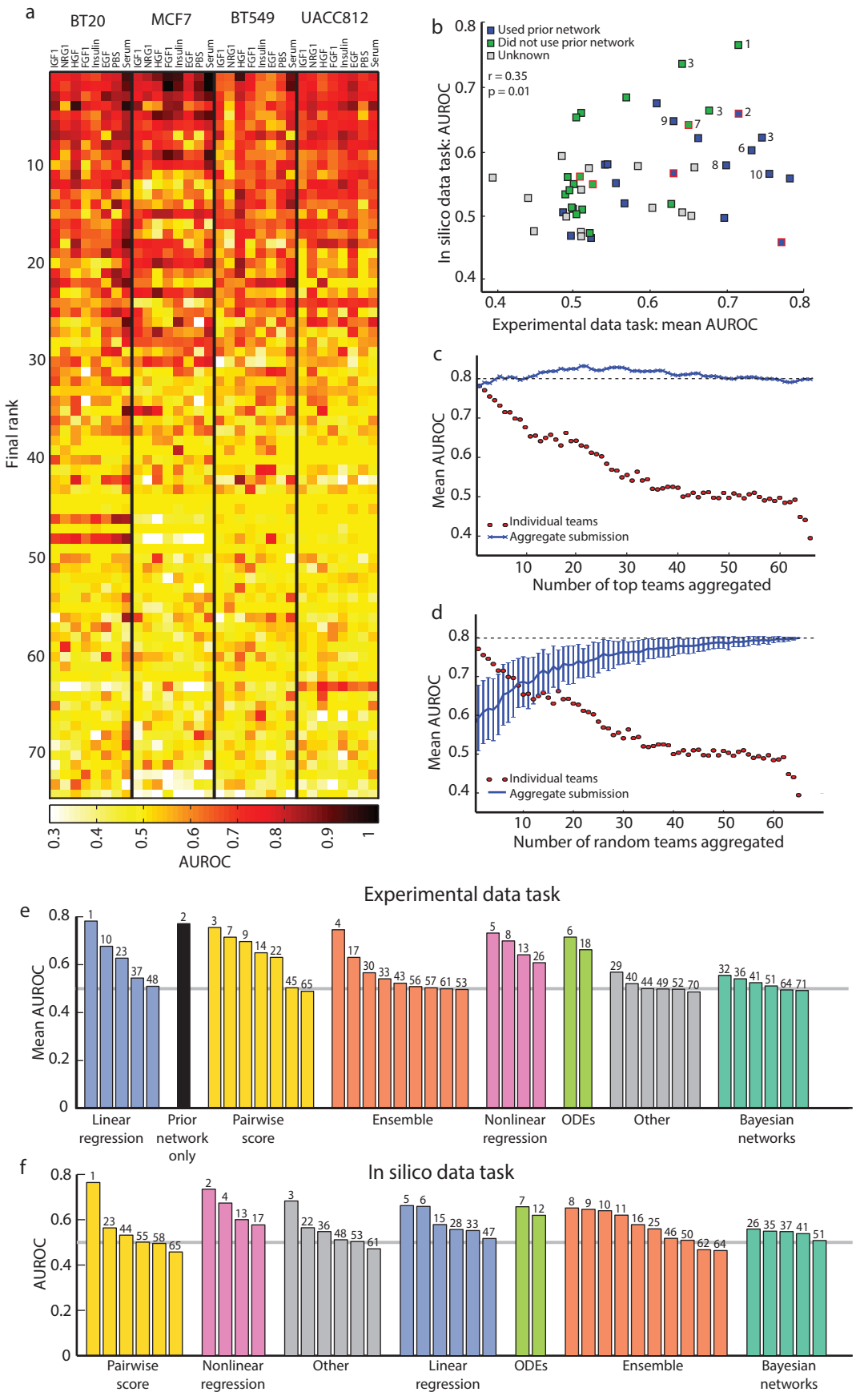
Figure-2 (Mukherjee)

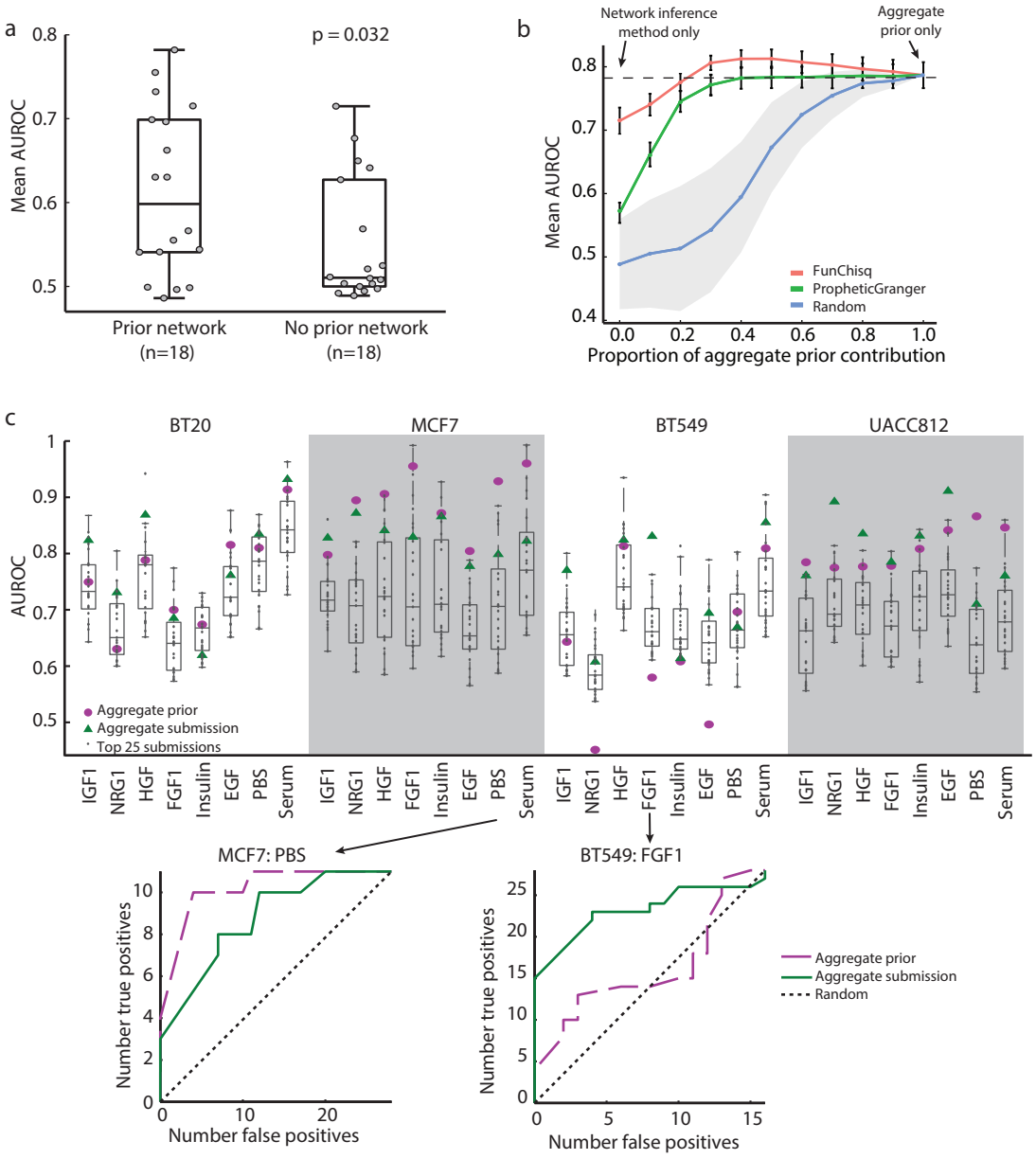


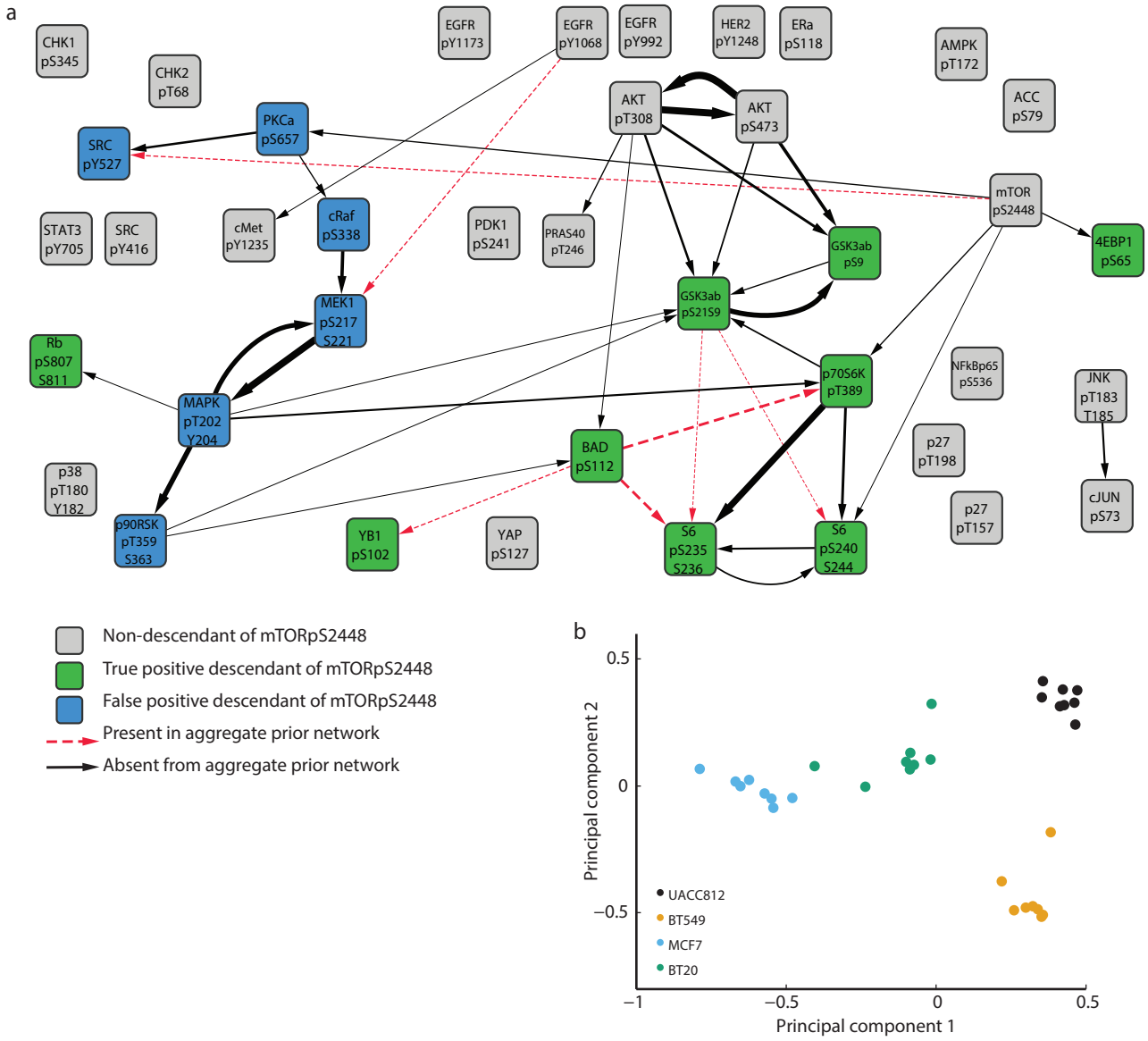







a
BT20
MCF7
BT549
UACC812
IGF1
NRG1
HGF
FGF1
Insulin
EGF
PBS
Serum
Final rank
10
20
30
40
50
60
70
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
AUROC
b
Used prior network
Did not use prior network
Unknown
r = 0.35
p = 0.01
In silico data task: AUROC
Experimental data task: mean AUROC
c
Mean AUROC
Individual teams
Aggregate submission
Number of top teams aggregated
d
Mean AUROC
Individual teams
Aggregate submission
Number of random teams aggregated
e
Experimental data task
Mean AUROC
Linear regression
Prior network only
Pairwise score
Ensemble
Nonlinear regression
ODEs
Other
Bayesian networks
f
In silico data task
AUROC
Pairwise score
Nonlinear regression
Other
Linear regression
ODEs
Ensemble
Bayesian networks


a
Mean AUROC
p = 0.032
Prior network
(n=18)
No prior network
(n=18)
b
Network inference
method only
Aggregate
prior only
Mean AUROC
FunChisq
PropheticGranger
Random
Proportion of aggregate prior contribution
c
BT20
MCF7
BT549
UACC812
AUROC
Aggregate prior
Aggregate submission
Top 25 submissions
IGF1
NRG1
HGF
FGF1
Insulin
EGF
PBS
Serum
MCF7: PBS
BT549: FGF1
Number true positives
Number false positives
Aggregate prior
Aggregate submission
Random


a
CHK1 pS345
CHK2 pT68
EGFR pY1173
EGFR pY1068
EGFR pY992
HER2 pY1248
ERa pS118
AMPK pT172
ACC pS79
AKT pT308
AKT pS473
PKCa pS657
SRC pY527
mTOR pS2448
4EBP1 pS65
STAT3 pY705
SRC pY416
cMet pY1235
cRaf pS338
PDK1 pS241
PRAS40 pT246
GSK3ab pS9
GSK3ab pS21S9
MEK1 pS217 S221
Rb pS807 S811
MAPK pT202 Y204
p70S6K pT389
NFkBp65 pS536
JNK pT183 T185
p38 pT180 Y182
BAD pS112
p27 pT198
cJUN pS73
p90RSK pT359 S363
YB1 pS102
YAP pS127
S6 pS235 S236
S6 pS240 S244
p27 pT157
Non-descendant of mTORpS2448
True positive descendant of mTORpS2448
False positive descendant of mTORpS2448
Present in aggregate prior network
Absent from aggregate prior network
b
0.5
0
−0.5
−1
−0.5
0
0.5
Principal component 2
Principal component 1
UACC812
BT549
MCF7
BT20