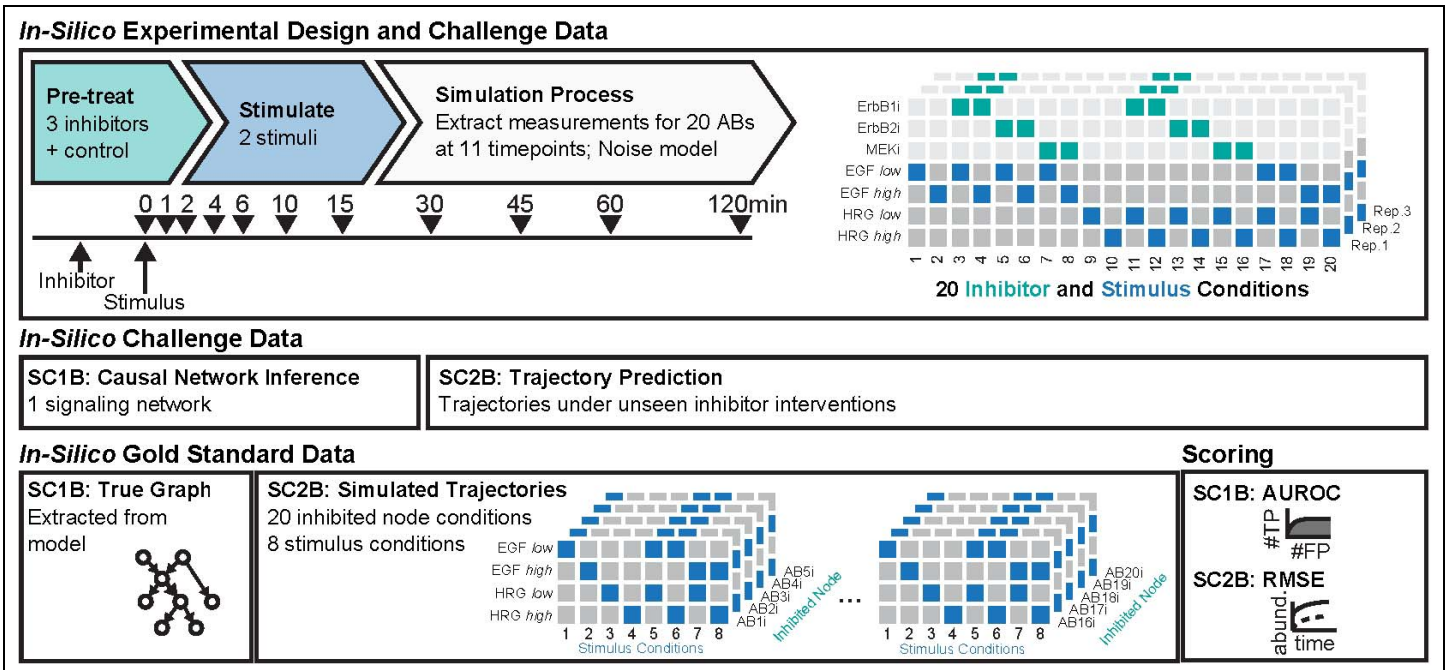
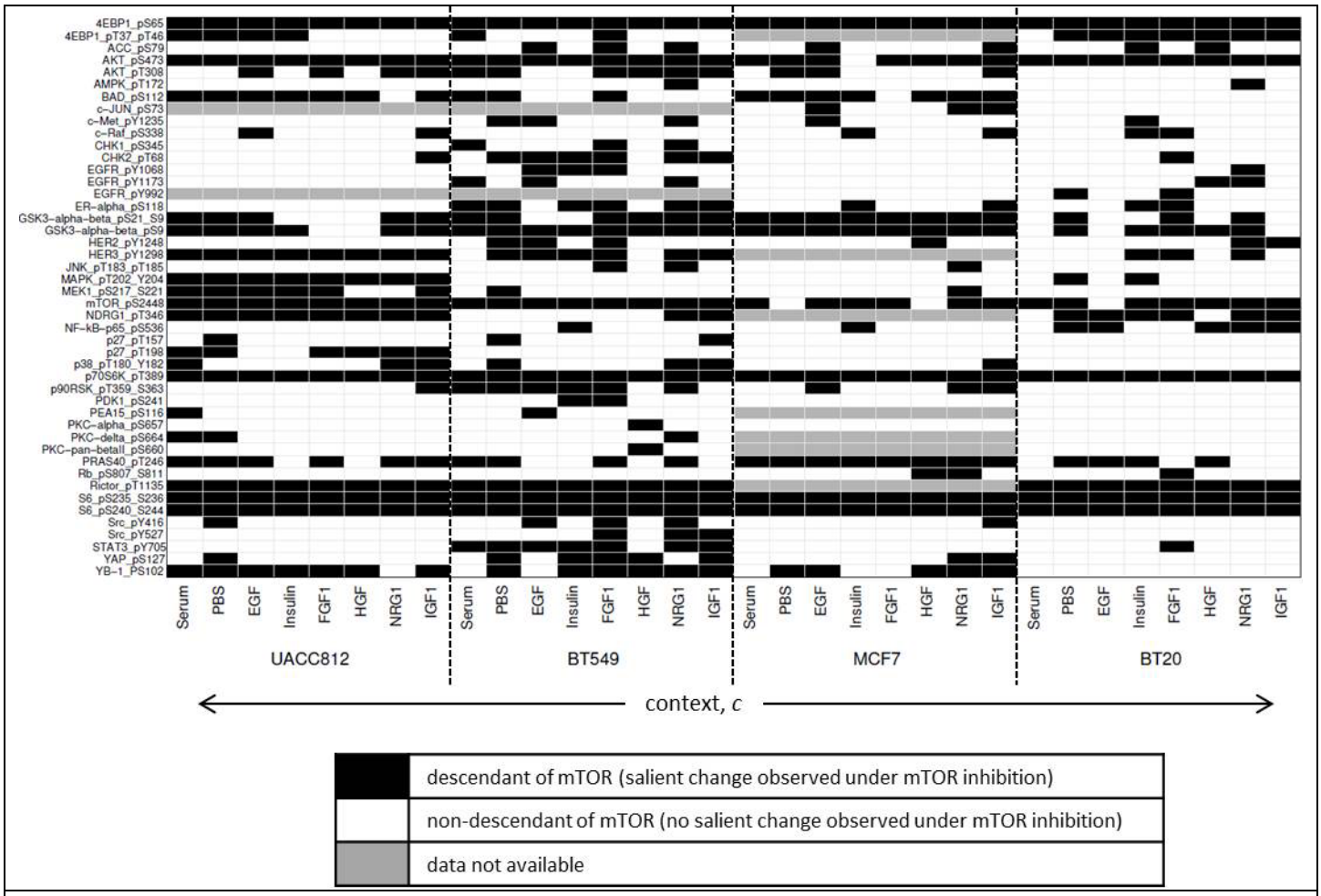


Supplementary Figure 1

The HPN-DREAM network inference challenge: overview of *in silico* data tasks.

Data were generated from a nonlinear dynamical model of the ErbB signaling pathway (Chen *et al*., 2009). Training data consisted of time-courses for 20 network nodes under three inhibitors targeting specific nodes, or no inhibitor, and under two ligand stimuli, applied individually and in combination at two concentrations. In total there were 20 different (inhibitor, stimulus) conditions as shown (top right). Time-courses comprised 11 time points and three technical replicates were provided. Node names were anonymized to prevent use of biological prior information. The sub-challenge 1 *in silico* data task (SC1B) asked participants to infer a single directed, weighted network using the training data. The aim of the sub-challenge 2 *in silico* data task (SC2B) was to predict stimulus-specific time-courses under unseen interventions. For SC1B, submissions were assessed against a gold-standard network extracted from the data-generating model, with agreement quantified using AUROC score. For SC2B, predicted time-courses were assessed using held-out test data obtained under *in silico* inhibition of each network node in turn, with prediction accuracy quantified using RMSE. See Online Methods for further details of the *in silico* data tasks.

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).

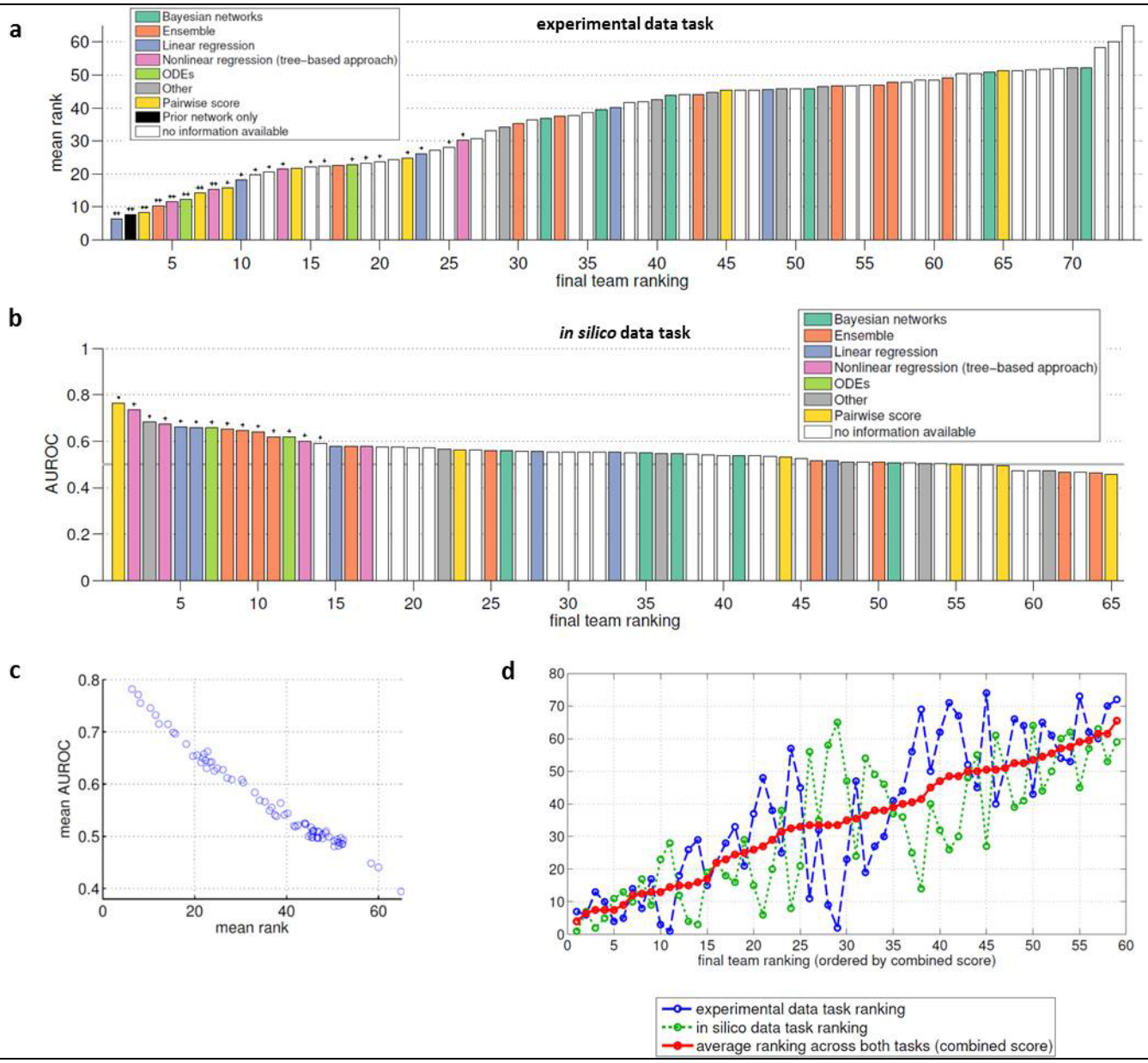


Supplementary Figure 2

Context-specific "gold-standard" causal descendant sets for the network inference sub-challenge experimental data task (SC1A).

Context-specific networks submitted to SC1A were assessed using held-out test data, obtained under inhibition of mTOR. Each column in the heatmap indicates, for a given *(cell line, stimulus)* context *c*, the phosphoproteins that showed salient changes under mTOR inhibition relative to DMSO control (black cells) and those that did not (white cells). Such changes were determined from the test data using a procedure centered around a paired *t*-test. Phosphoproteins that show salient changes can be thought of as descendants of mTOR in the underlying causal signaling network. Columns therefore represent context-specific sets of causal descendants of mTOR, D_c^{GS}, and were used as a "gold-standard" to assess inferred context-specific networks. Further details regarding the determination of the "gold-standard" descendant sets and the scoring procedure can be found in Online Methods. Missing data is indicated by gray cells (some phosphoprotein antibodies were only present in the (training and test) data for a subset of cell lines). Based on figure in Hill, Nesser *et al*. (Submitted).

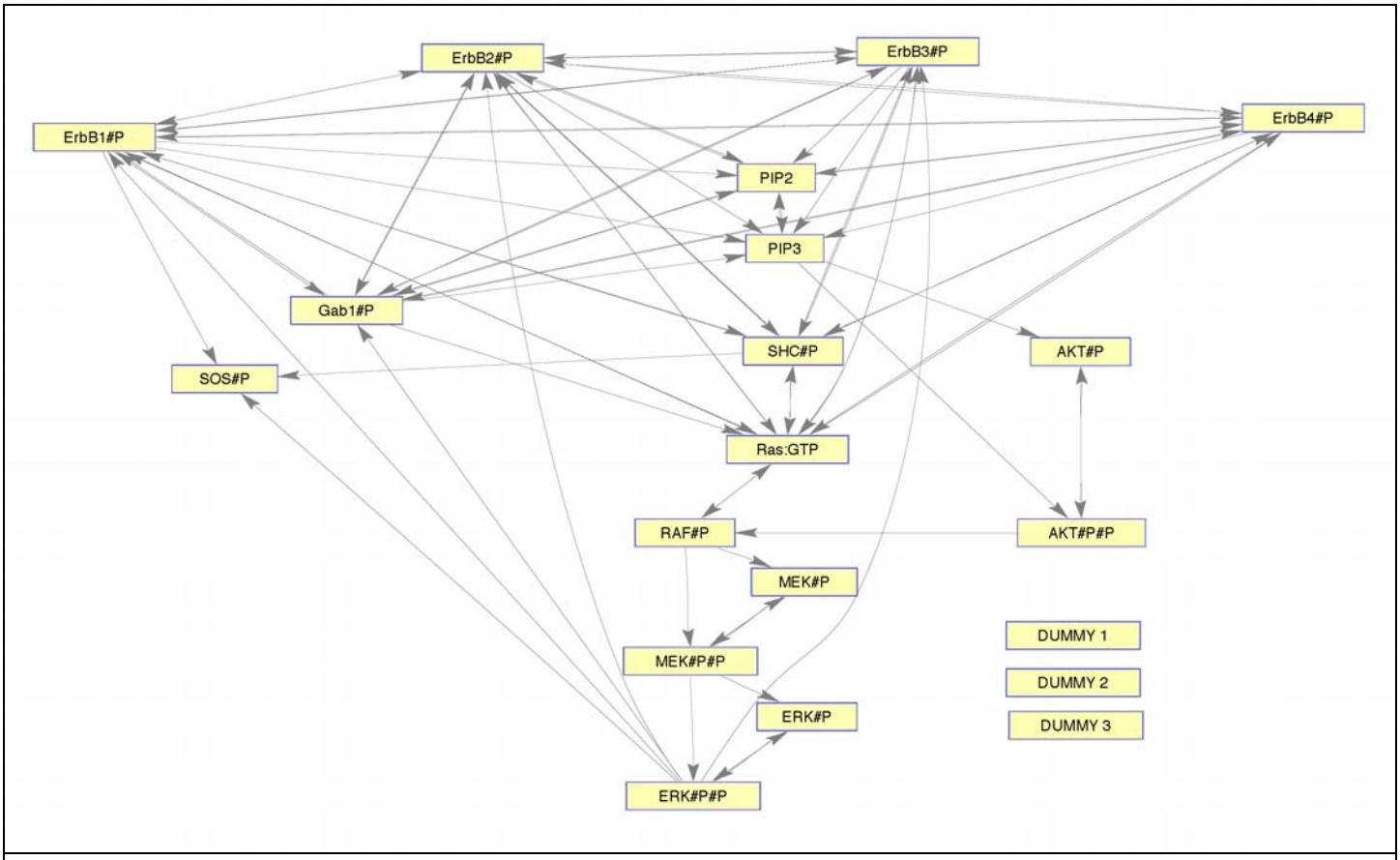
Hill, S.M., Nesser, N.K. *et al*. Context-specificity in causal signaling networks revealed by phosphoprotein profiling. (Submitted).



Supplementary Figure 3

Network inference sub-challenge (SC1) final team scores and rankings.

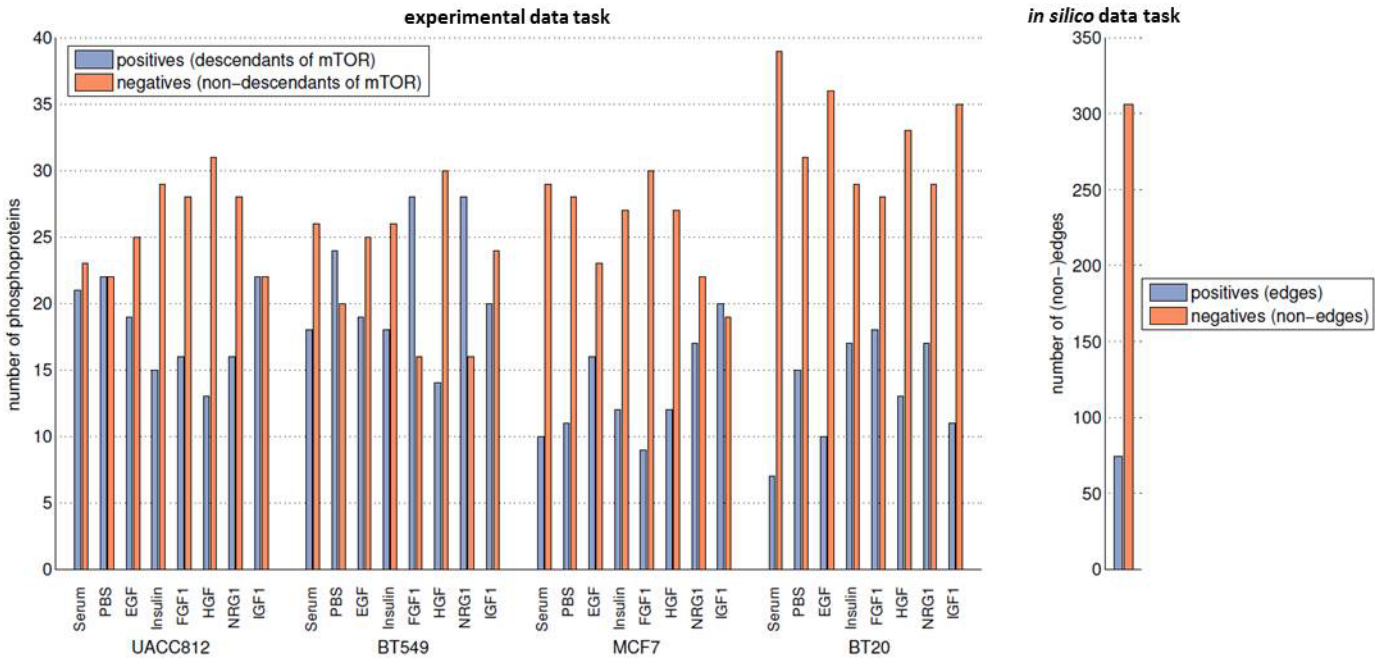
(**a**) Mean rank scores for the 74 teams that participated in the experimental data task (SC1A). Mean rank scores were used to obtain final team rankings. For the 40 teams that provided information regarding their approach, bar color indicates method type (see also **Fig. 3e**, **Table 1**, **Supplementary Table 2** and **Supplementary Note 5**). Stars above bars indicate teams with statistically significant AUROC scores (<5% FDR) in at least 50% of *(cell line, stimulus)* contexts (2 stars) or at least 25% of contexts (1 star). (**b**) AUROC scores for the 65 teams that participated in the *in silico* data task (SC1B). AUROC scores were used to obtain final team rankings. As in **a**, color indicates method type (see also **Fig. 3f**, **Table 1**, **Supplementary Table 2** and **Supplementary Note 5**). Stars above bars indicate statistically significant AUROC scores (<5% FDR). (**c**) Comparison of mean rank and mean AUROC scores for SC1A. Mean rank was used to obtain final rankings, but due to its strong correlation with mean AUROC, we used the latter in post-challenge analyses due to its interpretability. (**d**) Final ranks for SC1A (dashed blue line) and SC1B (dotted green line) were averaged to obtain a combined score (solid red line) for the 59 teams that participated in both tasks. Teams ordered by combined score (see "SC1A/B combined final rank" column in **Table 1** and **Supplementary Table 2**). See Online Methods for full details of scoring for SC1.



Supplementary Figure 4

Gold-standard causal network for the network inference sub-challenge *in silico* data task (SC1B).

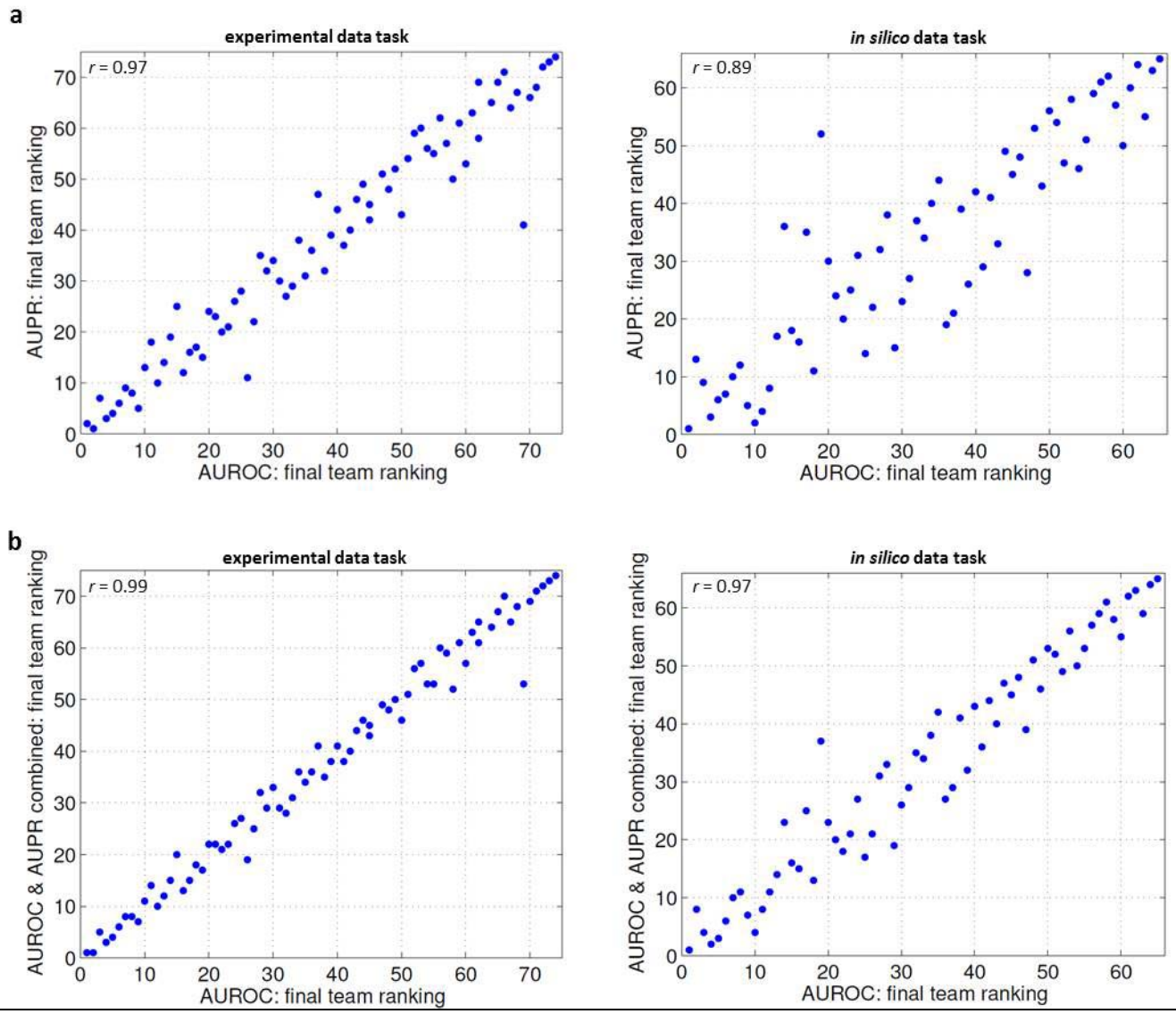
The gold-standard network, used to assess networks submitted to SC1B, was obtained from a data-generating dynamical model of the ErbB signaling pathway. Derivation of the network was non-trivial due to variables appearing in complexes within the model and full details can be found in **Supplementary Note 8**. Three unconnected dummy nodes were incorporated in the model and node names were anonymized in the training data.



Supplementary Figure 5

Balance of positives and negatives in the gold-standards for the network inference sub-challenge (SC1).

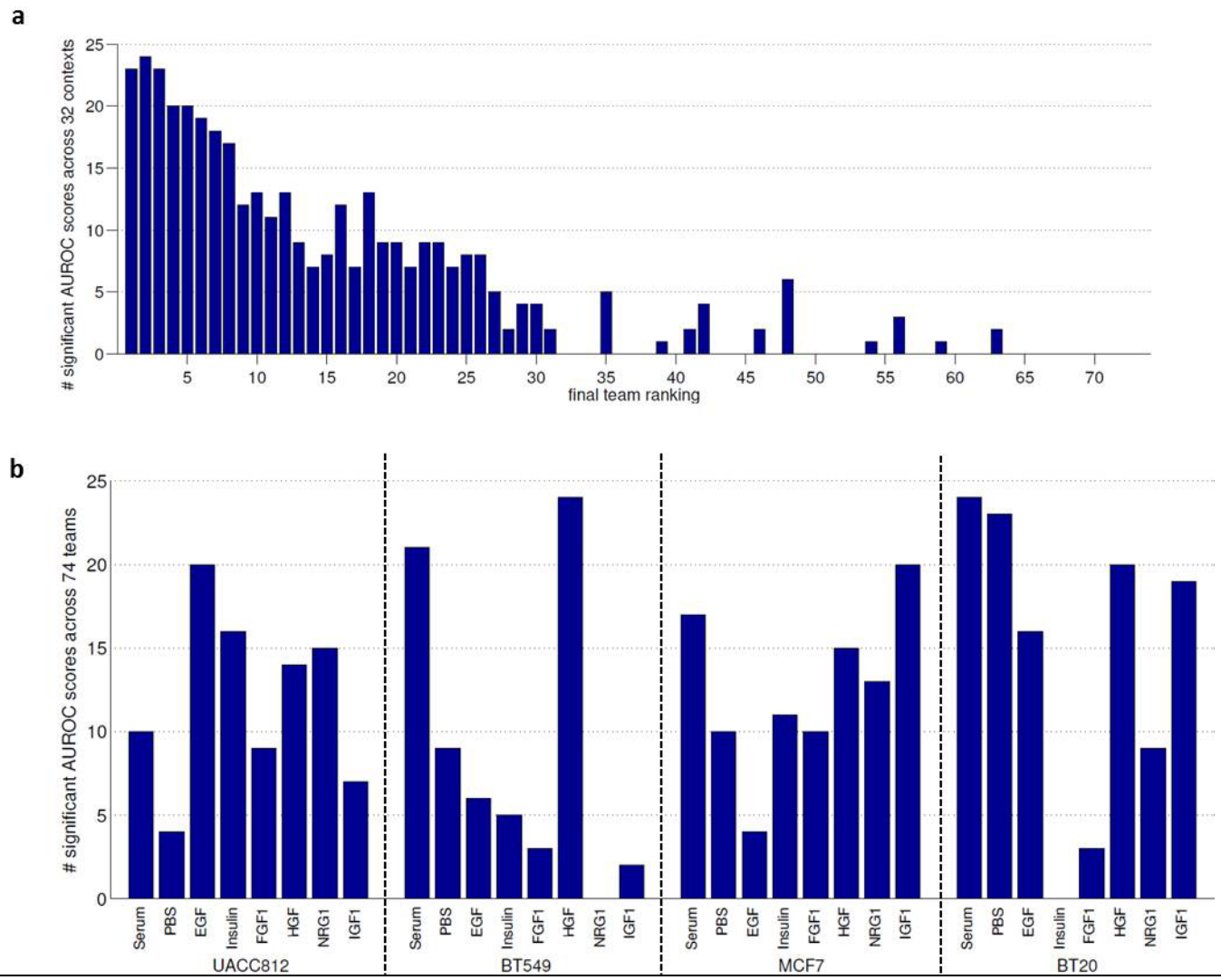
The “gold-standard” for the experimental data task (SC1A) comprised sets of descendants of mTOR for each *(cell line, stimulus)* context, determined from the held-out test data. Shown (left) are the number of positives and negatives in the “gold-standard” for each context; that is, the number of phosphoproteins that are descendants of mTOR according to the test data (positives) and the number that are non-descendants of mTOR (negatives). For the *in silico* data task (SC1B), the gold-standard consisted of the data-generating network. Shown (right) are the number of edges in this network (positives) and the number of non-edges (negatives).



Supplementary Figure 6

Comparison of AUROC with an alternative scoring metric, AUPR, for the network inference sub-challenge (SC1).

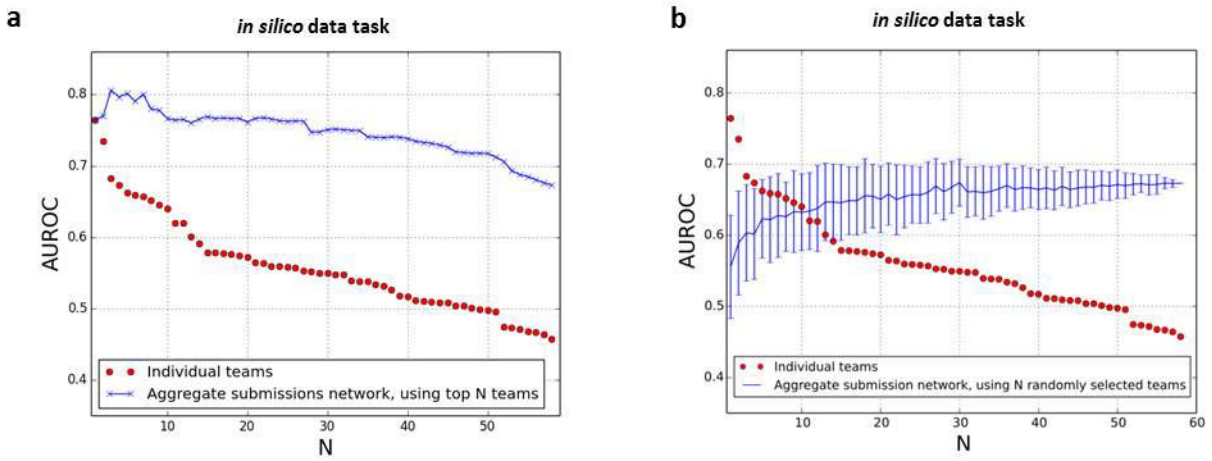
(**a**) Alternative team rankings were calculated by replacing AUROC with AUPR (area under the precision-recall curve) in the scoring procedure. The alternative rankings were compared with the original AUROC-based rankings for both the experimental data task (SC1A; left) and *in silico* data task (SC1B; right). (**b**) A further alternative ranking, combining both AUROC and AUPR, was obtained by ranking teams based on an average of final rank under AUROC and final rank under AUPR, and was compared with the original AUROC-based rankings.



Supplementary Figure 7

Statistical significance of AUROC scores for the network inference sub-challenge experimental data task (SC1A).

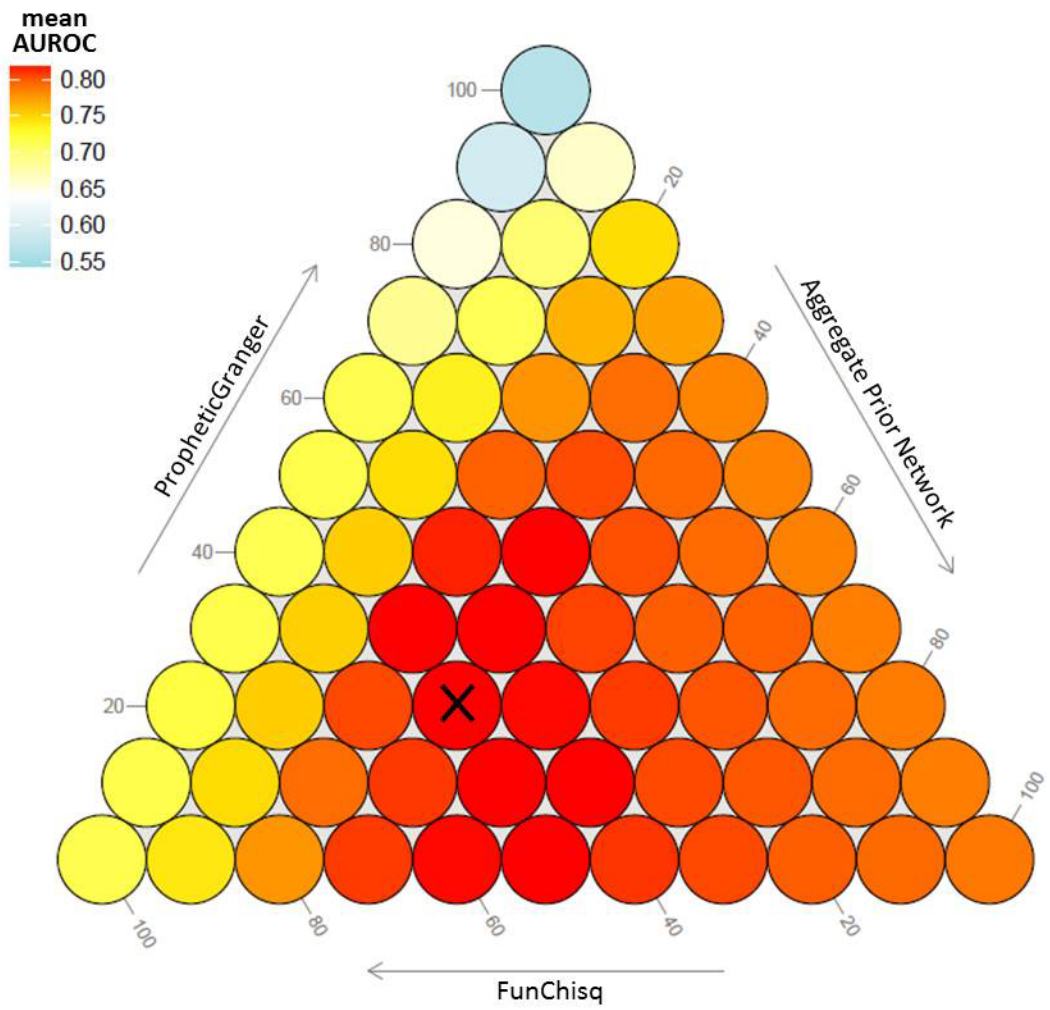
For each *(cell line, stimulus)* context, a null distribution over AUROC was generated and used to calculate an FDR-adjusted *p*-value for each team (Online Methods). (**a**) The number of significant (FDR<0.05) AUROC scores obtained by each team across the 32 contexts. Teams are ordered according to their final ranking in SC1A (based on mean rank score). (**b**) For each context, the number of teams (out of a total of 74) that obtained significant AUROC scores. For two regimes (BT549,NRG1 and BT20,Insulin), no teams obtained a significant AUROC score. These two regimes were disregarded in the scoring process.



Supplementary Figure 8

Crowdsourced analysis for the network inference sub-challenge *in silico* data task (SC1B).

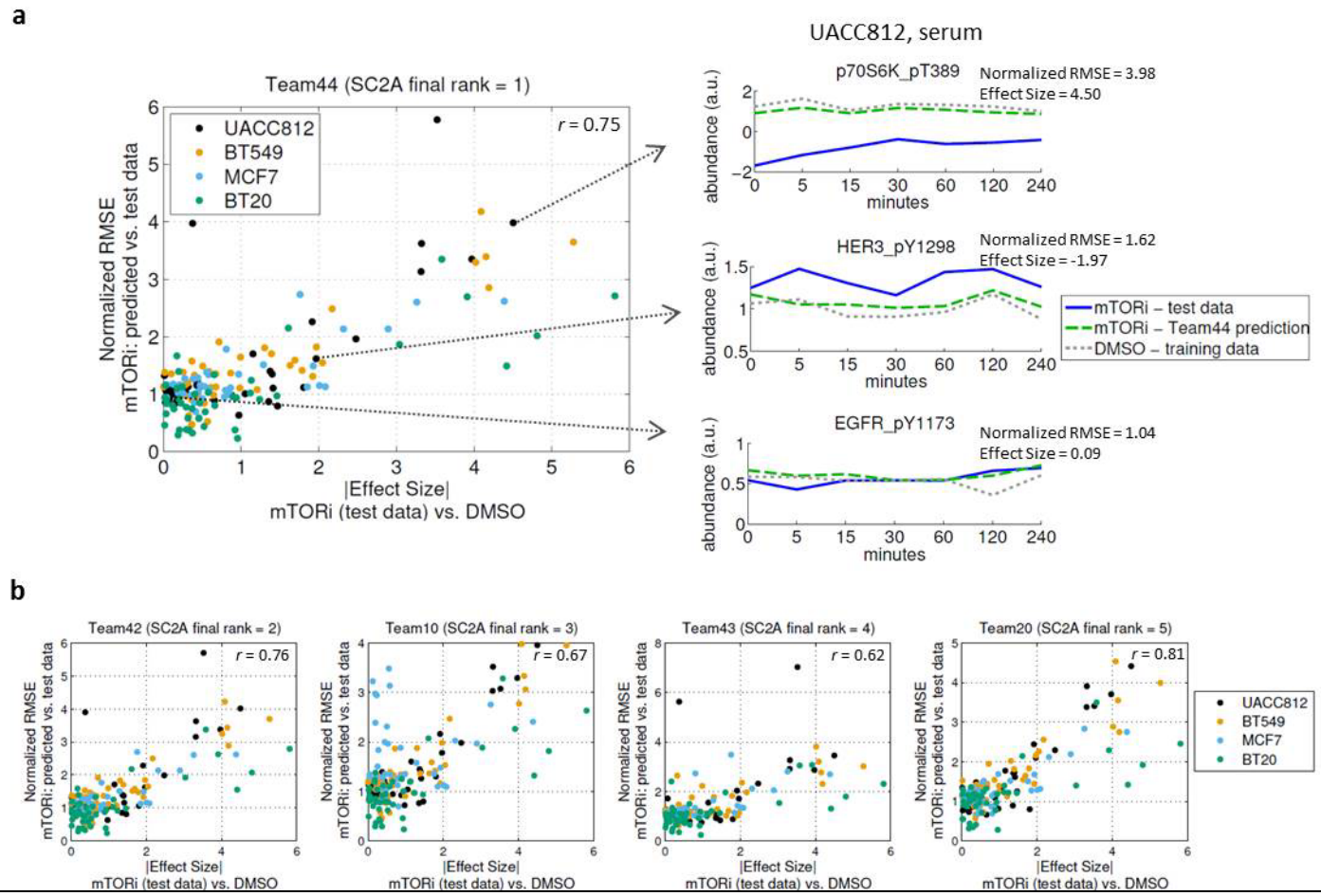
(**a**) Aggregate submission networks were formed by integrating predicted networks across the top *N* teams (as given by final team rankings), with *N* varied between 1 (top performer only) and all teams (after removal of correlated submissions; **Supplementary Note 10**). Integration was done by averaging predicted edge weights (Online Methods). The blue line shows performance (AUROC) of the aggregate submission networks with varying *N* . Individual team scores are also depicted (red circles). (**b**) Predicted networks were integrated for subsets of *N* teams, selected at random. The blue line shows mean performance of the aggregate submission networks with varying *N*, calculated over 100 random subsets of teams (error bars indicate standard deviation). Crowdsourced analysis for the experimental data network inference task is shown in **Figure 3c,d**.



Supplementary Figure 9

Weighted combinations of two top performing approaches and aggregate prior network for the network inference sub-challenge experimental data task (SC1A).

An extension of **Figure 4b** to show three-way combinations of (i) "PropheticGranger" – top performer for the experimental data task when combined with a prior network (here, the method is used without the prior network); (ii) "FunChisq" – top performer for the *in silico* data task and most consistent performer across both data types; and (iii) an aggregate prior network formed by integrating prior networks used by participants (Online Methods). The three approaches were combined by taking weighted averages of predicted edge scores for each *(cell line, stimulus)* context and performance assessed using mean AUROC. For example, the best performance (mean AUROC = 0.82) was achieved by combining 20% PropheticGranger, 50% FunChisq and 30% aggregate prior network, and is highlighted with an "X". See **Supplementary Note 1** for full details of the PropheticGranger and FunChisq approaches.



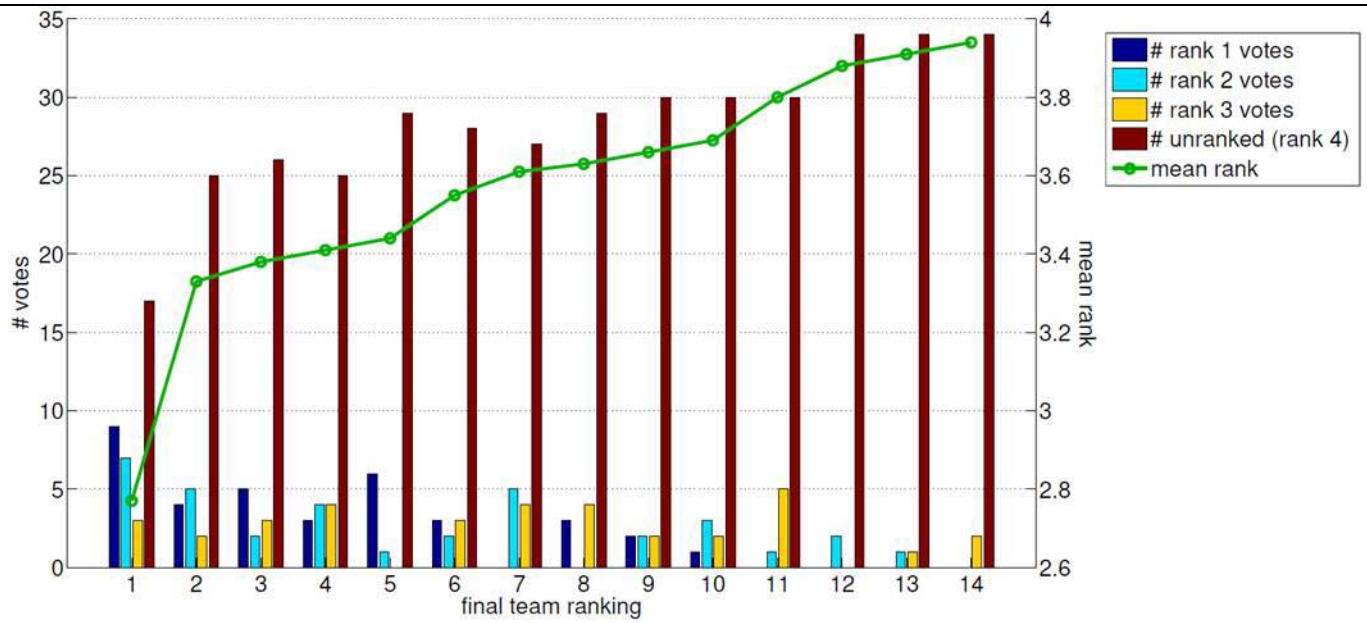
Supplementary Figure 10

Time-course prediction sub-challenge experimental data task (SC2A): phosphoproteins showing the largest changes under mTOR inhbition are predicted with least accuracy.

SC2A tasked participants with predicting phosphoprotein time-courses for each *(cell line, stimulus)* context under an unseen intervention (mTOR inhibition - mTORi). Submitted predictions were assessed against held-out test data obtained under mTORi. For each team, root mean squared error (RMSE) scores were calculated for each *(cell line, phosphoprotein)* pair (see **Supplementary Note 6**). (**a**) Left: for each *(cell line, phosphoprotein)* pair, normalized RMSE[1] for the top-ranked team (Team44) vs. absolute effect size. The effect size for a given *(cell line, phosphoprotein)* pair is a measure of the magnitude of abundance change under mTORi relative to DMSO control[2]. Note that this measure is based on the mTORi test data and is independent of team predictions. The strong positive correlation indicates that phosphoproteins showing little or no change under mTORi were predicted relatively well but phosphoproteins that showed large changes under mTORi were predicted badly. Right: examples of time-courses underlying the scatter plot (left). Shown are abundances of three phosphoproteins for cell line UACC812 under DMSO control and under mTORi, as predicted by Team44 and test data values. Note that normalized RMSE and effect size values are calculated across all stimuli, but only serum stimulus time-courses are shown here. (**b**) Scatter plots as in **a** for teams ranked 2 to 5 in SC2A. These results highlight the challenging nature of predicting protein abundance under unseen interventions but also point to a shortcoming of the RMSE score used here, namely that it does not sufficiently emphasize ability to predict proteins that change under intervention. For a future challenge, a modified metric that focuses on those proteins might therefore be useful.

[1]To ensure comparability across cell lines and phosphoproteins, each RMSE score was normalized by the standard deviation of the test data used in the RMSE calculation.

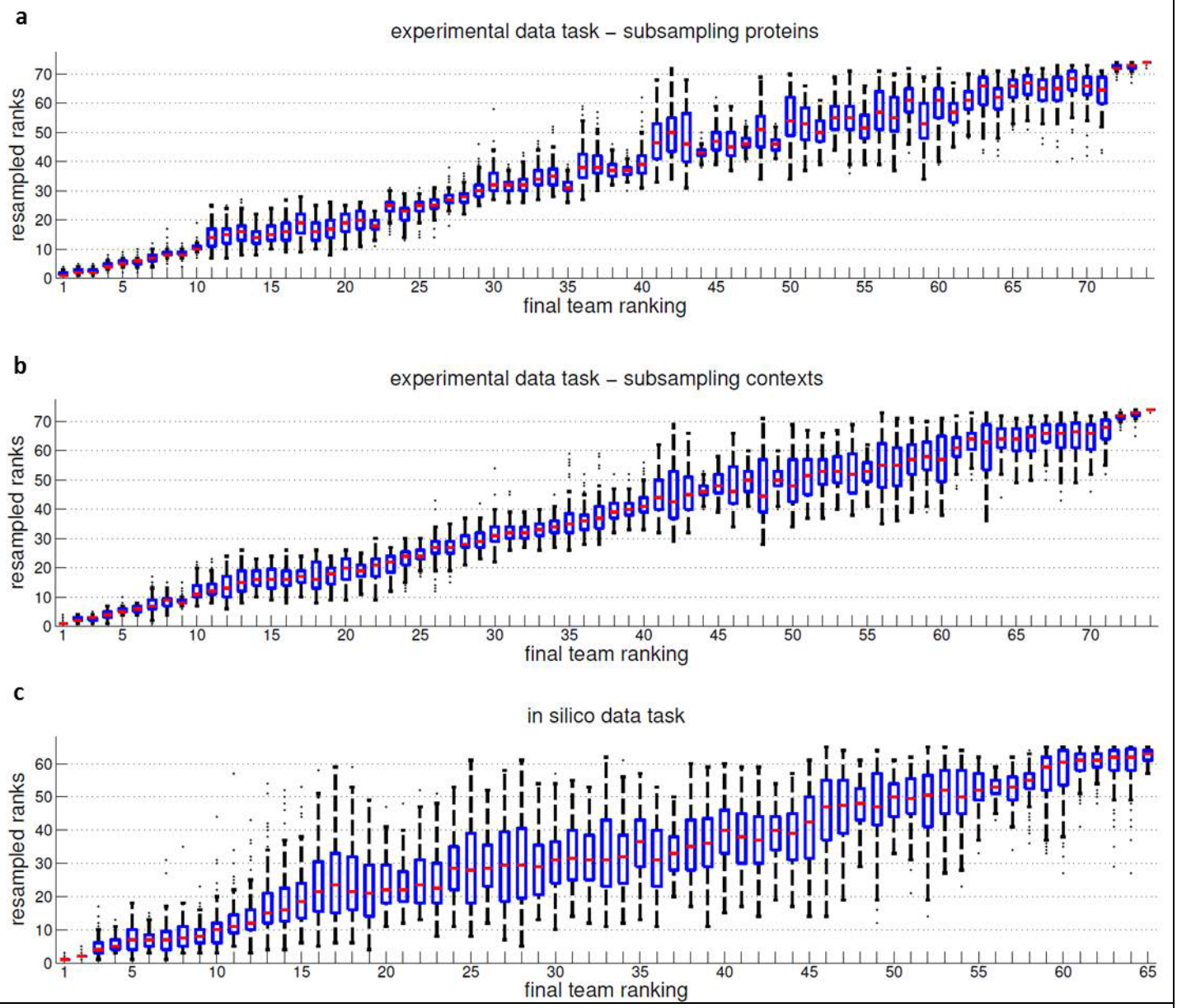
[2]Effect size is defined as the mean difference in phosphoprotein abundance between DMSO control and mTORi, normalized by the standard deviation of the differences. Means and standard deviations are calculated across all time points and stimuli for the given cell line.



Supplementary Figure 11

Visualization sub-challenge (SC3) voting results and rankings.

14 teams made submissions to the visualization sub-challenge. HPN-DREAM challenge participants were asked to select and rank (from 1 to 3) their three favorite submissions. The remaining unranked submissions were then assigned a rank of 4. 36 participants participated in the voting process and the number of votes of each rank type is shown (bar plot, left axis). Final team ranks were based on the mean rank across the 36 votes (green line, right axis).



Supplementary Figure 12

Robustness of rankings for the network inference sub-challenge (SC1).

The test data was subsampled to assess robustness of rankings (Online Methods). Tukey-style box plots show team ranks over 100 subsampling iterations, with 50% of the test data left out at each iteration. (**a**) Experimental data task - subsampling performed by removing 50% of phosphoproteins when assessing descendant sets for each *(cell line, stimulus)* context. (**b**) Experimental data task – subsampling performed by removing 50% of contexts from the scoring process. (**c**) *In silico* data task – subsampling performed by considering only 50% of edges/non-edges in the gold-standard network.