SCIENTIFIC REPORTS

Predicting Causal Relationships from Biological Data: Applying Automated Causal Discovery on Mass Cytometry Data of Human Immune Cells

Sofia Triantafillou [1,2], Vincenzo Lagani [1], Christina Heinze-Deml[3], Angelika Schmidt [4], Jesper Tegner[4,5] & Ioannis Tsamardinos [1]

Learning the causal relationships that define a molecular system allows us to predict how the system will respond to different interventions. Distinguishing causality from mere association typically requires randomized experiments. Methods for automated causal discovery from limited experiments exist, but have so far rarely been tested in systems biology applications. In this work, we apply state-of-the art causal discovery methods on a large collection of public mass cytometry data sets, measuring intra-cellular signaling proteins of the human immune system and their response to several perturbations. We show how different experimental conditions can be used to facilitate causal discovery, and apply two fundamental methods that produce context-specific causal predictions. Causal predictions were reproducible across independent data sets from two different studies, but often disagree with the KEGG pathway databases. Within this context, we discuss the caveats we need to overcome for automated causal discovery to become a part of the routine data analysis in systems biology.

Inferring causal relationships is of paramount importance in molecular biology. Knowing the causal structure of a molecular process allows advanced reasoning on its behavior, and such knowledge could be valuable in therapeutic approaches, for example through predicting side effects of pharmaceutical drugs. Given two factors A and B that affect each other, causal knowledge confers more information as compared to correlation. The correlation of factor A and factor B allows us to predict the levels of one given the levels of the other. However, it gives no information on the possible change in A if B is perturbed. This is not true for causal knowledge: Knowing the causal relationship of the two compounds allows predicting their response to an external intervention. For example, if A causally affects the abundance of B, then the perturbation of A is expected to affect the levels of B. Otherwise it will have no impact on B.

Computational causality has developed a language to describe, quantify and reason with causal claims. The most common framework of computational causality is causal Bayesian networks (CBNs), that use a simple assumption to connect causal relationships to associative patterns[1]. CBNs use directed acyclic causal graphs to describe the causal relationships and connect them to associations expected to hold or vanish in the joint probability distribution. Causal effects can also be computed using CBNs using do-calculus, a formal system for causal reasoning that includes an operation for interventions[1]. Algorithms for automatically identifying CBNs from limited or without experiments have also been proposed[2].

[1]Department of Computer Science, University of Crete, Rethimno, Greece. [2]Department of Physical Medicine and Rehabilitation, Northwestern University, Chicago, IL, USA. [3]Seminar for Statistics, ETH Zurich, Zurich, Switzerland. [4]Unit of Computational Medicine, Center for Molecular Medicine, Department of Medicine Solna, Karolinska Institutet & Karolinska University Hospital, Science for Life Laboratory, Stockholm, Sweden. [5]Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. Sofia Triantafillou and Vincenzo Lagani contributed equally to this work. Correspondence and requests for materials should be addressed to S.T. (email: sofi.triantafyllou@northwestern.edu)

The typical approach to learning causal relationships in current biology is by carrying out specifically designed experiments. Links that are established in the literature are then manually synthesized into larger causal models, like for example the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database[3]. Lately, high throughput techniques such as mass cytometry or single cell sequencing allow multivariate interrogation of large numbers of cells under a plethora of experimental conditions, with advanced technical reliability. Thus, applying techniques from the field of computational causality to systems biology could help revolutionize the time consuming, expensive and error-prone process of experimentally identifying causal structure in molecular systems.

In a seminal paper for applied causal discovery[4], the authors were able to almost flawlessly reconstruct a known causal signaling pathway from a mixture of experimental and observational flow cytometry measurements. This work illustrates the feasibility of causal discovery in biology. However, we must point out that several factors aided this success: the set of variables included in the analysis were not affected by any known latent confounders, and a mixture of observations and perturbations were included, facilitating the correct orientation of the recovered edges. The known system also included a limited number of feedback cycles (which were not identified correctly by the algorithm). *De novo* identification of causal relationships in data where latent confounders and feedback loops are possible, and interventions do not necessarily have known targets, can be far more challenging. A major contribution of this work is to elucidate the challenges of de novo identification of causal relationships.

In this work, we attempt to discover novel causal relationships from a large collection of public mass cytometry data of immune cells perturbed with a variety of compounds. Like flow cytometry, mass cytometry is a technique that can be used to singularize cells and measure protein abundance on the cellular level, resulting in very large sample sizes that are suitable for causal discovery methods. We discuss how different types of experiments can be modeled in the context of causal discovery, and then test the applicability of two state-of-the-art methods to identify phosphorylation of signaling proteins among the measured variables. We test the reproducibility of algorithmic results in similar, albeit different, data sets. Finally, we examine whether algorithmic findings are validated in the literature and in experimental data from the same study.

We find that (a) results are highly consistent on data sets that include different donors, experimental cell-stimulation time-points, or cell types (b) different causal methods often disagree with each other and with known pathways, (c) validity in experimental data is inconclusive. Our approach also revealed a large variation of the correlation structure, even in technical replicates produced within the same lab. Code for reproducing the results is available in https://github.com/mensxmachina/Mass-Cytometry. These results indicate that (a) de novo discovery of causal pathway relations is still a challenging task for current causal discovery methods, despite the previous positive results[4] and (b) current causal discovery methods do identify reproducible findings across similar data sets. There results constitute an important step for further developments in order to apply causal discovery methods successfully to biological single cell data in the future.

Results

Mass cytometry data.   As a test bed dataset to assess the applicability and performance of automated causal discovery methods, we used a public collection of mass cytometry data[5] (denoted 'BDM'). Multiplexed mass cytometry was used to measure the abundance of surface proteins and intracellular phosphorylated proteins in human peripheral blood mononuclear cells (PBMCs) that were stimulated with different factors (denoted 'activators') *in vitro*. The measured surface molecules are typically associated with certain PBMC subpopulations of specific function which, thus, can be distinguished by these markers. Within these cell populations, intracellular proteins specifically respond to certain extracellular signals to initiate regulatory pathways. To test these responses, known post-translational phosphorylations of specific proteins involved in immune cell signaling pathways were measured before and 30 minutes after treatment with several suitable activators.

The BDM study uses 11 different activators to stimulate cells, and furthermore it includes a multiple inhibitor experiment measuring protein responses in activated and non-activated cells under 8 dosages of 27 different inhibitors. For each inhibitor, the lower dosage is practically considered equal to zero, resulting in a collection of 27 replicates for the dose "zero". These data sets serve as a unique test bed for the reproducibility and consistency of algorithmic results. The measurement of the same system under slightly different experimental conditions should reflect similar mechanisms. However, it is worth noting that this is one of the first mass cytometry studies, and variation in the results could in part be attributed to lack of custom normalization techniques developed after its publication[6].

In the original publication, PBMC samples were manually gated into subpopulations based on the abundance of surface markers. The analysis is performed on this gated data, so each prediction is subpopulation-specific. Figure 1 illustrates the available data sets for each single subpopulation. Different subpopulations have different roles in the immune system and will have different responses to external cues. Moreover, the abundance of proteins within a cell greatly depends on cell size. Pooling data from different subpopulations together would create spurious relationships confounded by cell size. However, we assume that similar subpopulations should exhibit similar behavior.

Surface markers were used to discriminate subpopulation types and are not expected to react to the activator conditions, particularly within a narrow time frame like the one used in this study. We therefore only included the 14 functional proteins (measured as phosphorylated proteins whose phosphorylation status can change rapidly as response to activation) in all subsequent analyses (see Supplementary Table 1).

Learning causal relations using conditional independence tests.   Causal discovery makes assumptions on the nature of causality that connect the observable data properties (i.e., the joint probability distribution of the observed variables) to the underlying causal structure. The most popular causal assumption, famous for inspiring Bayesian networks, is the Causal Markov (CM) assumption. The CM assumption states that every variable is independent of its non-effects given its direct causes.
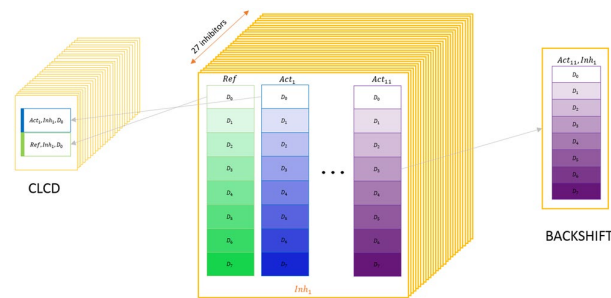
Figure 1. Collection of public mass cytometry data sets used for causal discovery. (middle) Available data for each subpopulation: single cells from a healthy human donor were selected and treated with 8 different dosages of a single inhibitor (D_0 corresponding to the zero dosage), and then stimulated with 11 different activators or without activator (denoted 'reference' data with no activation). The experimental plate setup with the different activations was repeated for 27 different inhibitors (with 8 doasages each). (left) Data usage for one run of CLCD: only zero dosages were used. Activated and non-activated data with no inhibition were pooled together to form a new data set which includes a binary column indicating activation status. The process was repeated for all 27 plates (inhibitors), and predictions were decided based on all 27 data sets. (right) Data usage for one run of BACKSHIFT: for every activator and inhibitor, different dosages of the same inhibitor are used as different environments.

In computational causality, the causal structure of a set of variables is often modeled directed graphs. Causal Bayesian Networks are the most popular framework for causality. In CBNs, causal relationships are modeled using directed acyclic graphs: A directed edge from X to Y denotes that X causes Y *directly* in the context of measured variables; no measured variable included in the model mediates the relationships. The causal structure is assumed to be acyclic. In addition, no pair of variables in the graph can have an unmeasured common cause. Several extensions try to relax these assumptions[7–9]. For simplicity, we present the theory of causal discovery using conditional independence tests using CBNs. We later discuss the effect of possible violations on our proposed approach.

Given the causal graph, one can easily identify direct causes and non-effects of a variable. The CM assumption connects a given causal graph with a set of conditional independencies (CIs). For example, the causal model shown in Fig. 2e1 implies that in the joint probability distribution of A, S and T, A must be independent of T given S (symb. $A \perp\!\!\!\perp T|S$), while all variables are pairwisely dependent. Essentially, this conditional independence denotes that the conditioning variable (S) *mediates* the relationship of the variables it renders independent: any information A carries about T goes through S. Therefore, given the value of S, A and T share no additional information about each other.

Figure 2a–d shows an example of a conditional independence pattern in mass cytometry data. Treating natural killer cells with PMA induces the phosphorylation levels of protein Zap70 (pZap70). Thus, stimulation with PMA and pZap70 are dependent: the probability distribution of pZap70 values given PMA stimulation is different than the marginal probability of pZap70 (Fig. 2b). However, given the levels of phosphorylation of protein Erk (pErk), stimulation with PMA is no longer informative for pZap70 (Fig. 2c; small differences in PMA-induced vs Reference pZap70 that are evident for small values of pErk are not statistically significant).

Conditional independencies can be tested in the data using appropriate tests of independence. Computational causal discovery tries to reverse engineer the causal graph using tests of independence: given a pattern of conditional independencies, a causal discovery algorithm typically tries to identify the causal graph that is associated to this pattern through the CM assumption. To do so efficiently, algorithms assume that all observed independencies in the data are a product of the causal structure, rather than being "accidental" or "fine-tuned" properties of the model parameters. This assumption is known as the Faithfulness assumption[2], and it is employed by most causal discovery algorithms. Even though random parameters typically lead to violations of faithfulness with probability zero[10], its validity has been debated[11–13].

Most of the times, CM and faithfulness do not uniquely associate a CI pattern with a single causal model. For example, all causal graphs in Supplementary Fig. 1 are associated with the (single) conditional independence $A \perp\!\!\!\perp T|S$. However, additional background knowledge can be used to narrow the space of possible models and enable novel causal discoveries. For example, knowing that A has no causes among the measured variables would rule out all other models apart but the causal model in Fig. 2e1, and allow a novel causal prediction: S must be a cause of T.

This example illustrates a simple case where background knowledge coupled with some of the assumptions of causal discovery (CM, faithfulness, no feedback, no confounders) indicate a non trivial causal relationship. More specifically, assuming that: (a) the causal structure of three variables *X*, *Y*, *Z* can be described using a directed acyclic graph and no pair of variables is confounded, (b) A is a known uncaused entity and (c) $A \not\!\perp\!\!\!\perp T$, $A \not\!\perp\!\!\!\perp S$, $S \not\!\perp\!\!\!\perp T$, $A \perp\!\!\!\perp T|S$, then we can infer that S causes T. The conditional independence pattern can be identified using appropriate statistical tests. This method, called Local Causal Discovery (LCD), has been proposed for automatically mining causal relationships in large data sets[14]. This simple structure has also been successfully exploited in genome-wide association studies (GWAS), where Mendelian randomization is used as a known uncaused entity[15,16].

LCD tries to identify this structure in data, testing for all pairwise dependencies and the conditional independence of A and T given S. For the causal structure $A \rightarrow S \rightarrow T$, two remaining conditional independencies are implied from the faithfulness assumption. To be more conservative, we propose (conservative local causal discovery) CLCD, an extension of LCD that tests for all 6 conditional (in) dependencies, and uses two separate
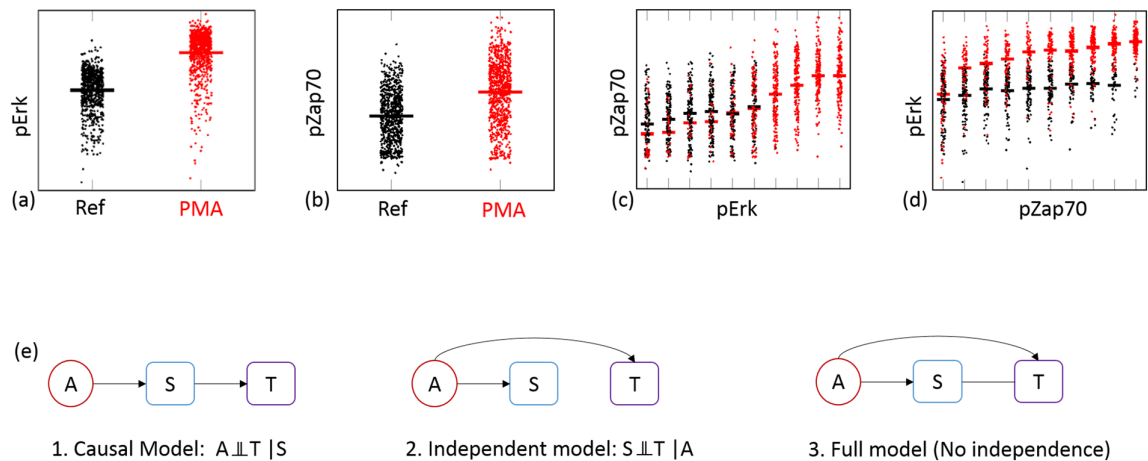
Figure 2. Causal discovery using conditional independence. (**a–d**) Joint and conditional distributions under conditional independence. Black/red points indicate no activation/activation with PMA. Black/red bars show the mean values. Conditional protein is discretized in 10 levels. (**a,b**) Stimulation with PMA significantly increases the levels of both pErk and pZap70. (**c**) pZap70 is independent of PMA stimulation given the levels of pErk: P(pZap70|PMA, pErk) = P(pZap70|pErk); mean values for red and black points are very similar. (**d**) In contrast, PMA stimulation is associated with pErk even after conditioning on pZap70. (**e**) Possible causal Bayesian networks for three variables that are pairwisely dependent, and one of them is restricted to be a source variable (no incoming edges): conditional independence can distinguish among the causal, independent and full model. In the full model, the line between variables S and T does not have any arrowheads; causal direction of this edge cannot be determined from the data.

thresholds, α and β for rejecting and accepting independence, respectively. Algorithm 1 describes the application of CLCD in the mass cytometry data.

Compared to trying to reconstruct the entire network of measured variables, considering only a small subset of variables at each run has many advantages for de novo causal discovery. Causal discovery algorithms are notoriously prone to error propagation[17], and a single error in a conditional independence test can affect seemingly remote parts of the learnt network. Employing a local approach allows testing all possible conditional independencies, and minimizes the probability that an independence test of unrelated variables will affect the output. Thus, error propagation does not affect the precision of CLCD.

We have so far presented CLCD assuming the "true" causal network is a CBN, i.e. there are no confounders and no feedback loops. However, violation of these assumptions does not have an effect on the validity of causal relationships identified with CLCD. Extensions of CBNs use bi-directed edges to model the presence of a confounder between two variables. Graphs that also include bi-directed edges are called mixed graphs. Conditional independencies implied by the CMC can be identified in mixed graphs using an extension of d-separation. the equivalent of d-separation for mixed graphs. The causal model shown in Fig. 2e1 is the only model that implies the conditional independence of A and T given S, for all possible networks, even in the presence of confounders (see Supplementary Fig. 2). In addition, the presence of a feedback cycle between S and T does not affect the validity of CLCD: In that case, the conditional independence of A and T given S no longer holds[8,9].

In mass cytometry experiments, the presence of an activator can be modeled as an external, binary variable, that is set by the experimenter and is not influenced by protein phosphorylation within the cell. All causal models where the activator is caused by a phosphorylated protein can then be excluded, and the conditional independence can only be explained if the mediation of S is causal. For example, the conditional independence of PMA and pZap70 given pErk supports that the correlation between pErk and pZap70 is a causal signal.

Notice that the data in Fig. 2a–d show the levels of pErk and pZap70 with and without PMA stimulation for the zero dosage of the Syki inhibitor. Similar data from the zero dosages of the remaining 26 inhibitors should agree with this CI pattern. Differences in the CI patterns in the 27 conceptually identical data sets could be accounted for by: (1) errors in the results of the conditional independence tests (2) effective inhibition even from a minuscule inhibitor dosage resulting in a different joint probability distribution.

Errors in the results of conditional independence tests can be either Type I errors (reject independence while it holds) or Type II errors (failing to reject an independence that does not hold). Type I errors can often be attributed to measurement noise, as measurement error in the mediating protein can result in failing to identify the conditional independence. Type II errors can be a result of weak dependencies and low sample sizes. In this work, we use two thresholds: a 0.001 threshold for rejecting independence, and a 0.15 threshold for accepting independence. This approach is more conservative than using a single threshold for accepting or rejecting independence. We also consider only triplets for which the same conditional independence pattern holds in at least 10 out of the 27 replicate data sets. We performed a sensitivity analysis for checking the effect of different thresholding on CLCD results (Supplementary material); this analysis indicates that different thresholding can affect the number of retrieved triplets, however does not modify the conclusions that can be drawn from the validations. Input data for this conservative local causal discovery (CLCD) pipeline are assembled from the original set of data as shown in Fig. 1.
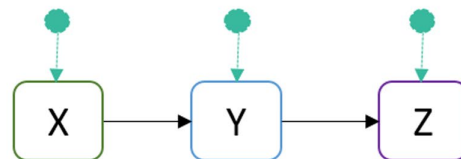
Environment 1 (observation)

$$X = \epsilon_x$$
$$Y = b_{xy}X + \epsilon_y$$
$$Z = b_{yz}Y + \epsilon_z$$
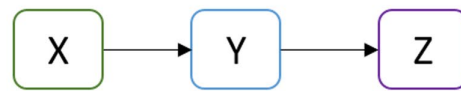
Environment 2 (intervention)
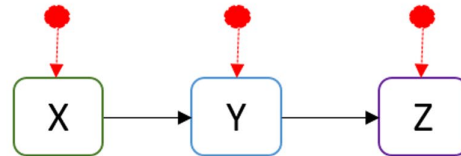
$$X = \epsilon_x + c_x^2$$
$$Y = b_{xy}X + \epsilon_y + c_y^2$$
$$Z = b_{yz}Y + \epsilon_z + c_z^2$$

Environment 3 (intervention)

$$X = \epsilon_x + c_x^3$$
$$Y = b_{xy}X + \epsilon_y + c_y^3$$
$$Z = b_{yz}Y + \epsilon_z + c_z^3$$

Figure 3. Causal discovery using different experimental conditions. A linear causal model is used to describe the causal structure of the measured variables. Different experimental conditions (environments) possibly "shift" the values of some of the variables (i.e. c_j^i can be zero for some i, j) while the causal structure of the variables remains invariant. BACKSHIFT exploits this invariance and can identify the causal structure given observations from at least three different environments.

Learning causal relationships using inhibitors as interventions with unknown targets. CLCD only uses a small portion of the available data, as it disregards all but the lowest inhibitor dosage. Data from different dosages of the same inhibitor correspond to different distributions and cannot be pooled together for CLCD. Some methods exist for integratively analyzing data from multiple experiments[18–20]. The experiments are required to be surgical interventions[18,19] (meaning that the targets of the experiments must be known and their values completely set by the experimental conditions). In the available data, different inhibitors can have various unknown targets, while different dosages of the same inhibitor change the magnitude of the effects. These types of interventions have been called "fat-hand" or "uncertain" interventions[20] or "shift interventions"[21].

In the recently developed method BACKSHIFT[21], different experimental conditions (or environments) with unknown targets are used to uncover causal structure. Different experimental conditions are modeled as so-called "shift interventions", meaning that the values of unknown targets of each intervention are shifted by realizations from a random variable modeling this intervention. Thus, the experiments are not required to be surgical interventions. In contrast, the underlying causal structure is assumed to persist in the experimental condition, affecting the system in addition to the induced intervention effect. The method then exploits the presence of different environments to identify the causal structure of the measured variables, as well as the location and strength of the interventions in each experiment.

To this end, the method assumes a linear causal model that possibly includes cycles and confounders. An example (without cycles nor confounders) is shown in Fig. 3. The coefficients b_{ij} fully describe the causal structure of the measured variables. Non-zero coefficients of the linear system correspond to direct causal relationships and their value corresponds to their strength. In Fig. 3, all b_{ij} coefficients apart from b_{xy} and b_{yz} are zero.

An intervention is modeled by the random variable c_j^i, where i denotes the experimental condition and j denotes the variable the intervention is acting on. The targets of the intervention do not have to be known and we do not require interventions to occur at every variable in each experimental conditions. In other words, some of the c_j^i can be zero.

The distribution of the noise terms as well as the coefficients b_{ij} are assumed to remain invariant across different experimental conditions. The estimation of b_{ij} then relies on a simple joint matrix factorization. At least three different environments, one of which can contain purely observational data, are required for identifiability. In this work, different dosages of the same inhibitor were used as different environments (i.e. dosages D_0 to D_7 shown in Fig. 1).

In the presence of various model violations, the joint diagonalization BACKSHIFT relies on is not possible. Therefore, model violations can be detected by the success or failure of the joint diagonalization algorithm. Inter alia, this applies to the following situations: (i) an intervention acted on hidden variables; (ii) interventions in the same environment are correlated; (iii) interventions act on the children of a certain variable which has more than one child. To increase robustness of the predictions, we used stability selection[22] which provides a finite sample control on the number of false discoveries.

As different inhibitor/activator combinations may behave in different ways, e.g. being specific or acting broadly, some data sets may fulfill the assumptions required by BACKSHIFT while others do not. Therefore, we considered only those results where the following conditions hold: (i) the joint diagonalization bootstrap confidence interval indicates that the joint diagonalization was successful; (ii) when refitting the model on subsamples of the data in the stability selection procedure, we require the joint diagonalization to be successful in at least 75 out of 100 runs. When these conditions were not met, we excluded the corresponding datasets from further consideration as the model assumptions of BACKSHIFT could be violated in these cases.

Predictions. Figure 4 shows the causal relationships predicted by CLCD and BACKSHIFT. Predictions are aggregated over all different activators (for CLCD and BACKSHIFT) and inhibitors (for BACKSHIFT). For the subpopulations not shown for BACKSHIFT in Fig. 4, the graph returned by the stability selection procedure was either empty, or obtaining the point estimate itself was not successful (joint diagonalization did not succeed).

Reproducibility in independent data sets. Apart from the data used in the CLCD and BACKSHIFT pipelines, the same study[5] includes time course data measuring protein response to several activators in PBMC samples across (a) eight different time points after activation ("time course data"), and (b) eight different donors ("8 donor data"), all measuring the same activators and proteins used in the main dataset with inhibitor treatments that was used above for CLCD and BACKSHIFT. Since BACKSHIFT demands different environments, it is not applicable in these additional data sets. In contrast, CLCD makes use of stimulated and unstimulated data, and the predictions should be consistent in matching conditions, meaning that a causal discovery algorithm that uses the same principles as CLCD should find the same causal network.

The exact CLCD pipelines are not applicable on the different time course and different donor data sets, so we used an (asymptotically equivalent) exact scoring scheme to test CLCD predictions, based on the BGEu score[23]. As a baseline, we estimated random predictions in a stratified manner (i.e. using the same number of predictions as those identified by CLCD for the each activator and subpopulation), over 10 iterations.

Predictions are significantly consistent in all time points (Fig. 5, p-values: 0.03, 10^{-34}, 10^{-10}, 10^{-56}, 10^{-34}, 10^{-52}, 10^{-9}, 10^{-7} for 0, 1, 5, 15, 30, 60, 120 and 240 minutes, respectively). For the first time points after activation (0 & 1 minute), sparser networks where activation is not yet effective or has not yet reached the target protein are selected for the majority of the predictions (Supplementary Fig. 3). Activation is effective for most of the predictions after 5 minutes, and the same causal model ($A \rightarrow S \rightarrow T$) is confirmed for the majority of the predictions at 30 minutes. Activation effects decline after thirty minutes, returning to sparser selected networks. The majority of the predictions are not contradicted in the time course data.

Similarly, the predictions are significantly consistent for all 8 healthy donors (Fig. 5, p-values: 10^{-19}, 10^{-5}, 10^{-13}, 10^{-4}, 10^{-65}, 10^{-54}, 10^{-6}, 10^{-8}). However, the actual percentage of consistent predictions is pretty low. A more careful inspection of the reproducibility results for the 8 donor data reveal that, in the majority of cases, protein levels are unaffected by the activation (see Supplementary Fig. 3), indicating possible problems in the experimental treatment.

To further test the reproducibility of CLCD predictions, we used not only the time course data and 8 donor data from the same study[5] but in addition, we analyzed a completely independent dataset from another study measuring surface proteins and intracellular protein phosphorylation levels in bone marrow immune cells of two healthy human donors, 15 minutes after stimulation with various activators[24] (denoted "bone marrow data"). Some of the proteins and activators are shared between the two studies (see Supplementary Table 1). While the studies are by no means identical, we tested whether similar phosphorylation patterns emerge in similar cell subpopulations upon activation with the same activators.

We used the same exact scoring scheme using the BGEu score[23] to test CLCD predictions. The bone marrow data are gated according to a different gating hierarchy, resulting in subpopulations that do not have a one-to-one correspondence to the PBMC data (Supplementary Table 2). Each subpopulation specific triplet was therefore tested in all similar populations (B-cells, T-cells and NK cells, monocytes, dendritic cells). Despite differences in cell type, gating hierarchy and activation time, almost half of the predictions are consistent for each donor. (Fig. 5, donor 1: $p = 10^{-9}$, donor 2: $p = 10^{-24}$). Moreover, for networks that were not confirmed in the bone marrow data, the highest-scoring networks do not contradict the prediction (Supplementary Fig. 3): for most inconsistent triplets, the full model is predicted, suggesting that the stimulation of the target protein could not be mediated by the target protein alone.

Validation of causal predictions. Without targeted randomized control trials, causal claims are hard to prove. While the published data include responses of the measured proteins to increasing dosages of 27 inhibitors, the inhibitions are typically not absolutely target-specific, and definitely not surgical. One inhibitor that can be considered an exception is Rapamycin, which is a very specific inhibitor known to target the mTORc1 protein complex, while the closely related complex mTORc2 is largely unaffected. The mTORc1 kinase complex is directly upstream of the S6K protein, which phosphorylates S6 (to obtain pS6) if the complex is active[25]. Data from experiments using the inhibitor Rapamycin, thus, seem to be suitable to test causal predictions. For example it could be tested whether a predicted downstream target of pS6 would be affected by the presence of Rapamycin.
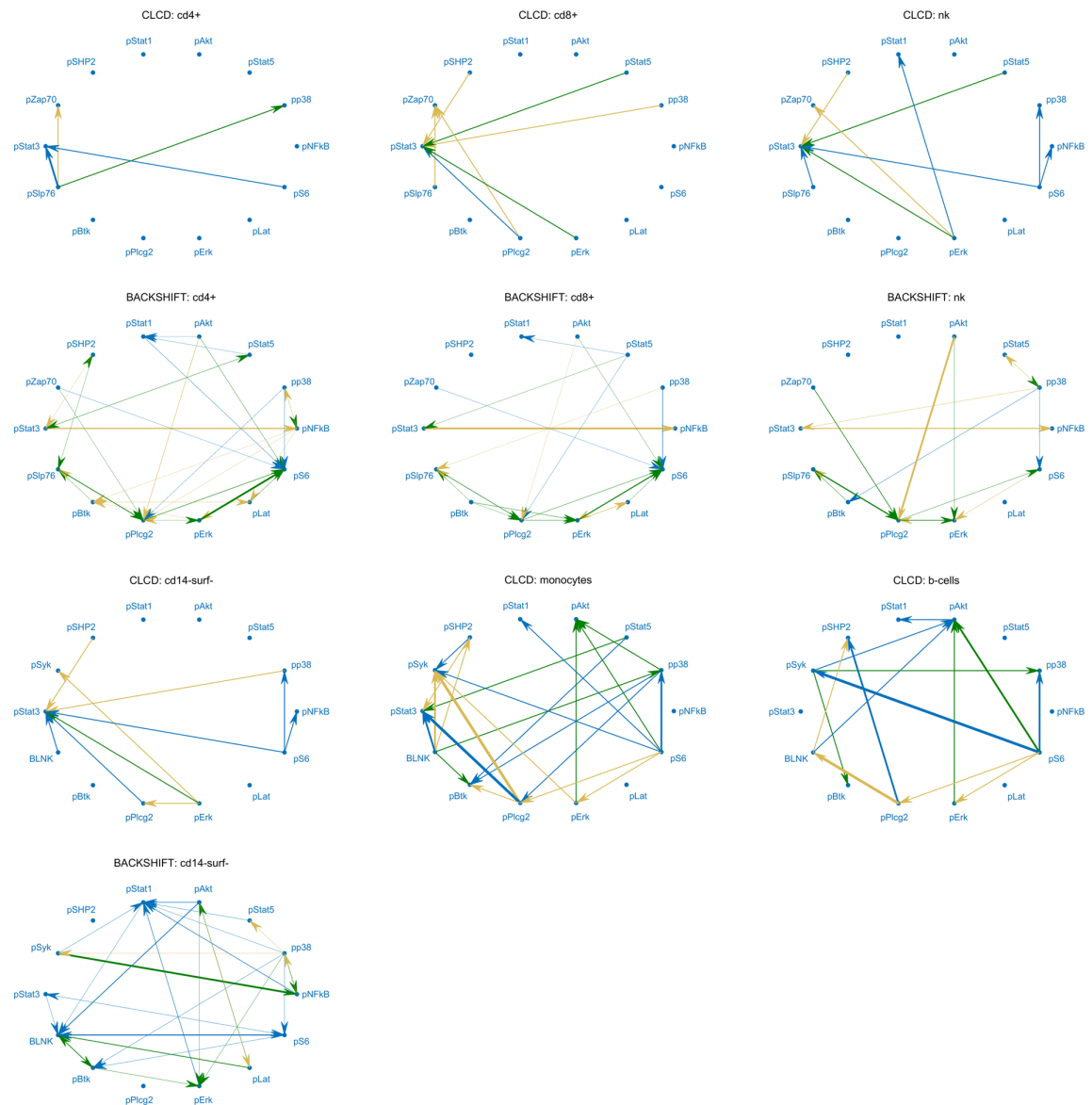
Figure 4. Predicted causal pairs for CLCD and BACKSHIFT. CLCD outputs 82 different predictions (38 unique cause-effect pairs), while BACKSHIFT outputs 193 predictions (51 unique cause-effect pairs). BACKSHIFT was successful for 4 out of 14 populations: CD4+ and CD8+T cells, Natural Killer (NK) cells, and CD14-surface- cells. CLCD results for the remaining subpopulations are grouped into CD14+ surface- and monocytes (CD14+HLA−Dr−, CD14+HLA−Dr^{high}, CD14+HLA−Dr^{mid}, CD14−HLA−Dr−, CD14−HLA−Dr^{high}, CD14−HLA−Dr^{mid}) and B cells (IgM− and IgM+). Edge thickness corresponds to frequency of appearance in different contexts (activators, inhibitors, subpopulations where applicable). Green edges are confirmed in at least one KEGG pathway, while brown edges are found reversed in at least one KEGG pathway. A detailed list with the predictions can be found in Supplementary Table 3.

Successful testing of such prediction requires the experimental system to show the expected behavior for known controls, that is, pS6 should decrease in the presence of Rapamycin. However, increasing doses of Rapamycin did not inhibit the phosphorylation of S6 in the public data we examined, as shown in the dose response curves for Rapamycin and its known target pS6, along with newly predicted targets of S6 in response to Rapamyin in Supplementary Figs 5 and 6.

    This indicates that either the cells in the experimental source data do not react as expected to respective stimulators and/or inhibitors, the chosen time points were not optimal for the respective readout, or the antibodies to detect the respective phosphoproteins did not work. The latter can be excluded however for pS6 since an inhibition of pS6 signal is seen in CD4+ T cells with the (unspecific) pan kinase inhibitor Staurosporine (see original publication[5] Fig. 4d). As also displayed in the original publication, it can be furthermore concluded that although
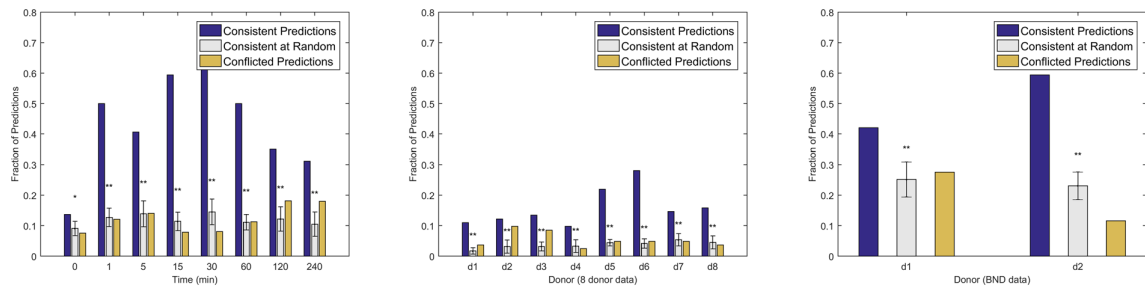
Figure 5. Reproducibility of the predictions in similar data sets. For every CLCD prediction ($A \rightarrow S \rightarrow T$), all possible networks were scored in an independent data set, and prediction is *consistent* if the same network is selected by the scoring scheme. In contrast, if the highest scoring network contradicts the predictions (i.e. includes a $S \leftarrow T$ edge), the prediction is considered conflicted. Predicted networks are found consistent in a variety of data sets, measuring the same variables and activators in (right) 8 distinct time points, (middle) 8 different healthy donors and (right) bone marrow data from two healthy donors in an independent study[24]. Asterisks denote statistical significance at 0.05(*) or 0.001(**). A one-tailed t-test is used to compare consistent predictions to consistency at random.
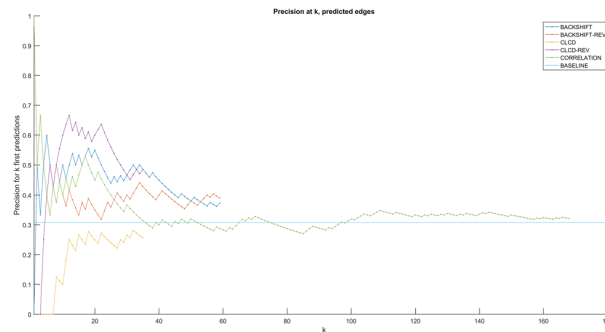
Figure 6. Precision of CLCD and BACKSHIFT predictions in KEGG human pathways. Cause-effect predictions were ranked according to frequency of appearance and then tested in KEGG human pathways. Out of 182 possible directed pairs, 56 pairs have a (direct or indirect) causal relationship in KEGG (baseline prediction). BACKSHIFT performed better than CLCD and mere correlation, but no method was significantly better than chance.

CD4+ and CD8+ T cells reacted with S6 phosphorylation to PMA/Iono stimulation, no inhibition of this pS6 signal was seen when the S6 inhibitor Rapamycin was applied (see original publication[5] Supplemental Fig. S19).

To compare our predictions to existing literature, we consulted the KEGG database which includes a collection of human protein signaling pathways based on literature from a multitude of different experimental systems. While the list of causal relationships documented in the pathways is not exhaustive, we examined the status of predicted relationships in the KEGG pathways. The 301 KEGG pathway entries corresponding to human pathways were downloaded and causal ancestry relationships among the 14 measured intracellular proteins were automatically mined using KEGGParser[26].

Figure 6 shows the precision of ranked predictions in the KEGG validation. Nine out of 38 CLCD unique predictions (regardless of activator and subpopulation) were identified in KEGG. We also tested the precision of the reverse relationships (i.e., reversing the causal direction of each prediction) in KEGG. Intriguingly, for 17 out of 38 predictions, the reverse relationship was identified in KEGG. While we do not have a clear explanation for this phenomenon, we hypothesize that this might be due at least in part to the presence of feedback cycles over time. BACKSHIFT achieved better performance: 22 out of 60 predictions were confirmed in KEGG pathways, while 24 out of 60 predictions were found reversed. We also compared our methods to predicting based on mere correlations, i.e. every correlated pair is used as causal in both directions. As a baseline for every method, we used 10 random predictions of the same size. Significance compared to chance was then assessed using a z-test. BACKSHIFT performs slightly better than correlation, however, neither of the two causal discovery methods was significantly better than random. However, notice that the comparison with the baseline is not very precise: the predictions were tested in *all* KEGG human pathways, regardless of the subpopulation and activator used to stimulate the cells. Since different cells can react with different pathways and to diverse stimulators based on the receptors, signaling components and regulators they express, this may contribute to mismatch of the pathway relationships even though some may still be true in reality.

Discussion

Biological data are often noisy, and highly dependent on the quality and availability of the used tools. In the analyzed data, a limitation of the CyTOF method is that the signal can only be as good as the available antibodies to detect the respective proteins, and the utilized inhibitors are not all absolutely specific for their desired target. Furthermore, the use of human primary cells naturally leads to donor variability with regard to response to stimulation, inhibition and expression of the measured proteins, as compared to for example measurements on more uniform cell lines. Another factor that complicates causal analyses is the frequent presence of (negative and positive) feedback loops in biological signaling systems.

Nevertheless, mass cytometry data have several characteristics that make them suitable for causal discovery: namely, they do not suffer from population averaging and come in adequately large sample sizes. Flow cytometry data, which are very similar in principle, have been used successfully in the past[4].

In this work, we examined the performance of two causal discovery methods on a collection of mass cytometry data. The methods are based on fundamental causal principles, and use multiple data sets and/or different experimental conditions to increase robustness. However, we found that even basic methods often disagree with each other and with background knowledge (such as the KEGG pathways). This effect might be due to violations of the assumptions of the two algorithms, although we found this point hard to demonstrate. On the other hand, results are shown to be reproducible in independent data sets, showing statistical patterns associated with certain causal models are consistent across different studies.

For automatic causal discovery, these results illustrate the need for (a) systematic study of the effect of violations of causal assumptions to existing algorithms, and (b) novel causal discovery methods that focus on conservative, quantitative, testable predictions in the presence of confounders and feedback cycles. Moreover, these results indicate the importance of practical applications for causal discovery algorithms; such interdisciplinary applications must be carefully selected, designed and tested in collaboration with domain experts, who can also guide the evaluation and extension of algorithms based on their domain-specific knowledge.

Methods

Gated data from[5] and[24] were downloaded from Cytobank (http://cytobank.org/). For every combination of inhibitor, activator (or absence thereof), subpopulation and dosage the number of samples ranges from less than 10 to a few hundreds, depending mainly on the subpopulation. Data sets with less than 20 data points were not used in our analysis. CLCD pipeline is shown in Algorithm 1, where $p(X, Y|\mathbf{Z})_D$ is the p-value for the conditional independence of X and Y given $\mathbf{Z}$ in data set D.

Input: Bodenmiller data over 27 Inhibitors, 14 Populations, 11 Activators plus no activator measuring 14 proteins; thresholds α, β, γ

for *Every population Pop* **do**
  for *Every activator A* **do**
    for *Every protein P_1* **do**
      for *Every protein $P_2 \neq P_1$* **do**
        $Score_{A \to P_1 \to P_2} \leftarrow 0$;
        for *Every inhibitor Inh* **do**
          Dataset $D \leftarrow$ Concatenate the zero dosage of the Inhibitor with Activator A the zero dosage of the Inhibitor with no Activator; add column indicating the presence or absence of Activator:;
          if $p(P_1, A|\emptyset)_D < \alpha$,
          $p(P_2, A|\emptyset)_D < \alpha$,
          $p(P_1, P_2|\emptyset)_D < \alpha$,
          $p(P_1, P_2|A)_D < \alpha$,
          $p(P_1, A|P_2)_D < \alpha$,
          $p(P_2, A|P_1)_D > \beta$ **then**
            $Score_{A \to P_1 \to P_2} ++$;
          end
        end
        if $Score_{A \to P_1 \to P_2} > \gamma$ **then**
          predict $A \to P_1 \to P_2$ in population *Pop*;
        end
      end
    end
  end
end

Algorithm 1. Application of CLCD on the Bodenmiller data

The following tests of independence are used:

- $P_1 \perp\!\!\!\perp A|\varnothing$, $P_2 \perp\!\!\!\perp A|\varnothing$: t-test.
- $P_1 \perp\!\!\!\perp P_2|\varnothing$, $P_1 \perp\!\!\!\perp P_2|A$: Fisher z-test[27].
- $P_1 \perp\!\!\!\perp A|P_2$, $P_2 \perp\!\!\!\perp A|P_1$: logistic test[28].

For BACKSHIFT, the causal system over p measurement variables $\mathbf{x}$ is defined by the set of linear relationships

$$\mathbf{x} = \mathbf{Bx} + \varepsilon, \tag{1}$$

where $\mathbf{B}_{i,j}$ is non-zero if x_j is a cause of x_i, and the error term ε is a p-dimensional random variable with mean 0 and positive semi-definite covariance matrix $\mathbf{\Sigma}_\varepsilon$.

The connectivity matrix $\mathbf{B}$ fully describes the causal structure of the measured variables: the non-zero coefficients of $\mathbf{B}$ correspond to the direct causal relationships, and the value of each coefficient describes the strength of the corresponding relationship. Allowing $\mathbf{\Sigma}_\varepsilon$ to be non-diagonal allows for possible latent confounders, and restricting the diagonal elements of $\mathbf{B}$ to be zero forbids self-loops. This is necessary for model identifiability.

For estimating the connectivity matrix $\mathbf{B}$, BACKSHIFT relies on observations of the system under shift interventions in different environments $e \in \mathcal{E}$

$$\mathbf{x}^e = \mathbf{Bx}^e + \mathbf{c}^e + \varepsilon, \tag{2}$$

where $\mathbf{c}^e \in \mathbb{R}^p$ are uncorrelated interventions (the covariance matrix of $\mathbf{c}^e$ is a diagonal matrix). Thus, the intervention is allowed to perturb each variable, while the causal relationships between the variables remain the same. Identifying matrix $\mathbf{B}$ is equivalent to fully identifying both the causal structure (non-zero elements) and the strength of causal relationships.

Estimation of $\mathbf{B}$ is based on the relation between the covariance matrices (i) $\mathbf{\Sigma}_\mathbf{x}^e = \text{Cov}(\mathbf{x}^e)$ – the covariance of $\mathbf{x}$ in environment e (observed), (ii) $\mathbf{\Sigma}_\mathbf{c}^e = \text{Cov}(\mathbf{c}^e)$ – the covariance of $\mathbf{c}$ in environment e (unobserved) and (iii) $\mathbf{\Sigma}_\varepsilon = \text{Cov}(\varepsilon)$ – the covariance of the noise (unobserved).

Exploiting the structure of the model, BACKSHIFT finds $\mathbf{B} \in \mathbb{R}^{p \times p}$ such that $\forall \ e \in \mathcal{E}$

$$(\mathbf{I} - \mathbf{B}) \underbrace{\mathbf{\Sigma}_\mathbf{x}^e}_{\substack{\text{function of } e, \\ \text{non}-\text{diagonal}}} (\mathbf{I} - \mathbf{B})^T = \underbrace{\mathbf{\Sigma}_\varepsilon}_{\substack{\text{invariant over } e, \\ \text{non}-\text{diagonal}}} + \underbrace{\mathbf{\Sigma}_\mathbf{c}^e}_{\substack{\text{function of } e, \\ \text{diagonal}}}.$$

The estimation relies on a simple joint matrix diagonalization, applied to differences between covariance matrices, and a reformulation of the linear sum assignment problem. This amounts to a computational complexity of $O(np^2)$ where n is the total number of observations across all environments. For further details about the method, we refer to[21]. Failure of the joint diagonalization algorithm indicates some model violation. The success or failure of the joint diagonalization algorithm is assessed by a parametric bootstrap approach.

For the stability selection, the parameters were chosen as $\mathbb{E}(V) = 5$, $n_{sim} = 100$ and $\pi_{thr} = 0.75$. This implies that the estimated network only contains those edges that were found more than 75 times when refitting the model on 100 random subsamples of the data. Thus, stability selection allows us to assess whether BACKSHIFT yields consistent estimates when different subsamples of the same dataset are used. Moreover, only these stable results are included in the final output. In the analysis, the expected number V of falsely selected variables was set to 5. In general, decreasing $\mathbb{E}(V)$ leads to sparser graphs.

References

1. Pearl, J. *Causality* (Cambridge university press, 2000).
2. Spirtes, P., Glymour, C. N. & Scheines, R. *Causation, prediction, and search* (MIT press, 2000).
3. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research* **44**, D457–D462 (2016).
4. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. & Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–9 (2005).
5. Bodenmiller, B. *et al*. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nature biotechnology* **30**, 858–867 (2012).
6. Finck, R. *et al*. Normalization of mass cytometry data with bead standards. *Cytometry Part A* **83**, 483–494 (2013).
7. Richardson, T. & Spirtes, P. Ancestral graph markov models. *Annals of Statistics* 962–1030 (2002).
8. Glymour, C. N. & Cooper, G. F. *Computation, causation, and discovery* (Aaai Press, 1999).
9. Richardson, T. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, 454–461 (Morgan Kaufmann Publishers Inc., 1996).
10. Meek, Ch. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (UAI'95)*, 411–418 (1995).
11. Robins, J., Scheines, R., Spirtes, P. & Wasserman, L. Uniform consistency in causal inference. *Biometrika* **90**, 491–515 (2003).
12. Uhler, C., Raskutti, G., Bühlmann, P. & Yu, B. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics* **41**, 436–463 (2013).
13. Mooij, J. M. & Cremers, J. An empirical study of one of the simplest causal prediction algorithms. In *UAI 2015 Workshop on Advances in Causal Inference*, 30–39 (2015).
14. Cooper, G. F. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery* **1**, 203–224 (1997).
15. Schadt, E. E. *et al*. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics* **37**, 710–717 (2005).
16. Chen, L. S., Emmert-Streib, F. & Storey, J. D. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome biology* **8**, R219 (2007).
17. Spirtes, P. Introduction to causal inference. *Journal of Machine Learning Research* **11**, 1643–1662 (2010).
18. Hyttinen, A., Eberhardt, F. & Järvisalo, M. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI 2014*, 340–349 (2014).
19. Triantafillou, S. & Tsamardinos, I. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *JMLR* **16**, 2147–2205 (2015).
20. Eaton, D. & Murphy, K. Exact Bayesian structure learning from uncertain interventions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 107–114 (2007).

21. Rothenhäusler, D., Heinze, C., Peters, J. & Meinshausen, N. BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems* **28**, 1513–1521 (2015).
22. Meinshausen, N. & Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society, Series B* **72**, 417–473 (2010).
23. Geiger, D. & Heckerman, D. Learning gaussian networks. In *UAI*, 235–243 (1994).
24. Bendall, S. C. *et al.* Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science* **332**, 687–696 (2011).
25. Magnuson, B., Ekim, B. & Fingar, D. C. Regulation and function of ribosomal protein s6 kinase (s6k) within mtor signalling networks. *Biochemical Journal* **441**, 1–21 (2012).
26. Arakelyan, A. & Nersisyan, L. KEGGParser: parsing and editing KEGG pathway maps in Matlab. *Bioinformatics (Oxford, England)* **29**, 518–9 (2013).
27. Fisher, R. The distribution of the partial correlation coefficient. *Metron* **3**, 329–332 (1924).
28. Lagani, V., Kortas, G. & Tsamardinos, I. Biomarker signature identification in "omics" data with multi-class outcome. *Computational and structural biotechnology journal* **6**, 1–7 (2013).

Acknowledgements

Author Contributions

S.T., V.L. and I.T. conceived the idea. S.T. and V.L. preprocessed the data. S.T., V.L. and C.H.D. analyzed the data. S.T., V.L. and A.S. designed the validation analysis. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-08582-x

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.