Consensus Enhances Individual Causal Models: a Use Case on Lung Cancer Driven by Cellular Pathways

Arnaud Lang[1][0009−0007−9115−5428], Rodrigo Henrique Ramos[2][0000−0002−0786−5387], Safaa Al-Ali[3][0000−0003−1864−8190], Mohammad Reza Mousavi[4][0000−0002−4869−6794], Anna Calissano[5][0000−0002−7403−0531], and Irene Balelli[1][0000−0002−4593−8217]

[1] Inria Center at Université Côte d'Azur, Epione Team, Sophia Antipolis, France {arnaud.lang,irene.balelli}@inria.fr
[2] Federal Institute of São Paulo, São Carlos - São Paulo, Brazil ramos@ifsp.edu.br
[3] Univ. Bordeaux, CNRS, Inria (Carmen Team), Bordeaux INP, IMB, UMR 5251, IHU Liryc, F-33400 Talence, France safaa.al-ali@inria.fr
[4] King's College London, Strand, London WC2R 2LS, UK mohammad.mousavi@kcl.ac.uk
[5] University College London, London, UK a.calissano@ucl.ac.uk

Abstract. Discovering reliable cause-and-effect relationships in real-world medical data is an open challenge. Classical Causal Discovery (CD) algorithms used to solve this task rely on strict assumptions that are rarely met in complex real-world scenarios with limited expert knowledge - the functional form of the causal relationships, the data distribution, the causal sufficiency. Thus, the reliability of CD algorithms can significantly drop, compromising the interpretability of the results and the trustworthiness of downstream decision-making. To overcome these limitations, we introduce the concept of *consensus causal model* to combine various CD algorithms and enhance their accuracy. Our consensus model can be efficiently constructed from a set of heterogeneous causal graph objects through a homogenisation step, ensuring semantic compatibility with the original edge definitions and enabling meaningful information exchange. To showcase the proposed method, we analyze a lung cancer dataset combining patient-level information such as smoking habits and age, and we study their effect on the onset and development of the disease, the tumor stage, and cellular pathway mutations. By applying multiple classical CD algorithms, we observe significant structural inconsistencies and heterogeneity across individual graphs. We demonstrate that the consensus causal model, unlike the individual models, effectively aggregates the strengths of each algorithm while mitigating their uncertainties. The resulting model reveals biologically validated causal relationships between risk factors, mutations, and pathways that isolated algorithms fail to capture, thereby underscoring the value of consensus causal modelling as a robust alternative to single-model selection for causal discovery.

Keywords: Causal Discovery, Consensus Causal Models, Benchmark, Knowledge-based systems, Cancer Genomics, Cancer Patient Data, Cellular Pathways.

1 Introduction

The growing complexity and the mostly opaque nature of artificial intelligence models have been raising concerns about their interpretability, explainability, and trustworthiness. Consequently, there is increasing interest in causality, a field that aims to move beyond correlations and uncover the mechanisms underlying data generation, thus providing enhanced predictive power and enabling reasoning about interventions and counterfactuals. This makes causality particularly relevant in critical domains such as e-healthcare [33,20], where potentially high-risk questions should be reliably answered, while caring for trust-building between the system (computational models for clinical decision-making support) and multiple users (clinicians, nurses, physicians, and importantly, patients). While controlled experiments such as randomized clinical trials remain the gold standard for establishing causality, they are often costly, time-consuming, or even infeasible due to ethical issues. This has motivated the development of computational approaches to infer causal relationships across the retained features of interest directly from observational data, by combining domain knowledge and data-extracted insights: *Causal Discovery* (CD) [17].

Reliably performing CD in real-world settings turns out to be an extremely challenging task [10]. Indeed, the theoretical identifiability of CD algorithms (*i.e.*, the guarantee of recovering the *true* underlying causal graph given sizeable, good-quality data) is tightly related to a strict compliance of the data with some difficult-to-achieve assumptions. Those assumptions can span the type and shape of the causal relationships between variables, the nature and additivity of the independent variable-specific error term, or the observation of an exhaustive set of causal variables (*cf.* causal sufficiency, see Sec. 2.1). Moreover, real-world (e.g., healthcare) datasets are often incomplete, noisy, or biased, further affecting practical identifiability.

In the absence of a robust strategy for assessing whether a dataset satisfies the specific assumptions required as a minimum to guarantee the theoretical identifiability of a given CD algorithm, it can be dangerous to rely on its results to draw conclusions, particularly in critical areas such as health. The simultaneous application of multiple CD strategies could reveal more consistent patterns and help researchers formulate more informed and cautious interpretations. In this work, we propose to move from an individual to a multiple models approach by aggregating the causal information yielded from several CD algorithms, rather than arbitrarily selecting a single model as the best representative one. The driving motivation is that we do not dispose of enough expert knowledge to decide on one unique model. Therefore, we introduce the concept of a *consensus causal model*, which will be obtained by first mapping CD outputs to a universal causal graph representation, facilitating the sharing of information, and then merging these homogenized causal graphs, each of them providing its partial yet valuable view of the same reality. The consensus causal model is expected to convey a richer and more informative structure, while naturally enabling a heuristic measure of the associated confidence, and overall improving trustworthiness.

We evaluate our approach on a high-impact use case: lung cancer. Cancer is a complex disease that emerges from the intricate interplay of mutations accumulated over time that corrupt cell pathways. Factors such as smoking and age can influence the emergence of the disease and contribute to its heterogeneous progression. The increas-

ing availability of (public) cancer data enables the development of many statistical studies exploring, for instance, the relationships between mutations, cellular pathways, and risk factors. CD can enhance these findings by moving from correlational to causal associations, thus bringing crucial insights for designing tailored treatments. In a recent paper, [34] proposed to analyse data from lung cancer, including patient data, tumor stage, driver genes, and cellular pathways, and apply the classical Peter-Clark (PC) algorithm [41] to discover causal associations from the patient to the genomic level. This allowed to reinforce confidence in some already known causal associations between patient and tumor data, while suggesting new potential causal relationships. Nevertheless, the choice of PC as the backbone of their pipeline for CD on lung cancer has not been discussed, and this is not an isolated example (*e.g.*, [24,22]). In this work, we consider the dataset analyzed in [34]. We demonstrate that applying several classical CD methods generates a strong heterogeneity across the resulting causal graphs, emphasizing the risks of relying on a single method when expert knowledge is not available. Certain causal relationships are more consistently discovered than others, and a clear cause-and-effect directionality can sometimes be extremely hard to establish. This motivates a frequency-based aggregation approach, in which edges are weighted by their observation frequency, yielding a more robust and informative estimate of the underlying causal structure.

The paper is organised as follows. In Section 2, we provide a brief introduction to causality, describe the CD algorithms included in our method, and present our method to build the consensus causal model. We also include a description of the dataset on lung cancer that we will consider in this study. Section 3 first presents the analysis of the observed heterogeneity within the set of discovered causal graphs obtained by running our CD benchmark. Then, we build our consensus model and provide a biological validation of the consensus graph, highlighting the relevance of our method as a valid and low-cost strategy for obtaining more comprehensive and accurate insights in our use case. Section 4 concludes the paper with some suggestions for future work.

2 Materials and Methods

2.1 Background on Probabilistic Causality

Given two random variables (rvs) X_i, X_j, a common-sense definition of causality can be stated as follows: if X_i causes X_j, then an intervention on X_i will affect X_j, whereas the inverse does not hold. Considering a set of rvs $\mathbf{X}$, a convenient way of representing causal relationships consists of a directed graph $\mathcal{G} = (\mathbf{X}, E)$, where E contains all directed edges relating nodes in $\mathbf{X}$: if X_i causes X_j, then $\mathcal{G}$ will contain a directed edge $X_i \to X_j$, *i.e.*, $(X_i, X_j) \in E$ (the order matters), we say that X_i is a parent of X_j and denote Pa_j the set of direct causes of X_j. We assume that a causal graph is *acyclic*, meaning that there exists no set of consecutive edges in E starting and ending in the same node: in this case, we talk of a Directed Acyclic Graph (DAG) (see [32] for an overview on the topic). The indirected graph associated with $\mathcal{G}$ is called the *skeleton*. Causal discovery aims to uncover, from data, the DAG underlying the causal data generation mechanism. Let $\mathbf{X}$ be a set of causal variables (sometimes called *endogenous*) and $\mathbb{P}$ a probability distribution over $\mathbf{X}$. Learning the causal DAG over $\mathbf{X}$ means

learning how the data sampled from the distribution $\mathbb{P}$ has been generated. We can connect probability and causality by looking at the (in)dependencies and conditional (in)dependencies in $\mathbb{P}$. The *Causal Markov Condition* allows to make a link between the causal graph and the probability distribution through conditional independence.

Definition 1. *Let $\mathcal{G} = (\mathbf{X}, E)$ be a causal graph, and $\mathbb{P}$ a probability distribution over $\mathbf{X}$ generated by $\mathcal{G}$. $\mathcal{G}$ and $\mathbb{P}$ satisfy the **Causal Markov Condition (CMC)** iff for every X_i in $\mathbf{X}$, $X_i \perp\!\!\!\perp \mathbf{X} \setminus (Des_i \cup \mathrm{Pa}_i) | \mathrm{Pa}_i$ (read: X_i is independent of $\mathbf{X} \setminus (\mathrm{Des}_i \cup \mathrm{Pa}_i)$ given Pa_i).*

The Markov Condition entails the following factorization of the joint probability distribution :

$$\mathbb{P}(\mathbf{X}) = \prod_{X_i \in \mathbf{X}} \mathbb{P}(X_i | \mathrm{Pa}_i). \tag{1}$$

We also recall the definition of *Faithfulness*, required by most CD algorithms for identifiability:

Definition 2. *Let $\mathcal{G} = (\mathbf{X}, E)$ be a causal graph, and $\mathbb{P}$ a probability distribution over $\mathbf{X}$ generated by $\mathcal{G}$. $\mathcal{G}$ and $\mathbb{P}$ satisfy the **Faithfulness Condition** iff the only conditional independencies in $\mathbb{P}$ are the ones entailed by the CMC applied to $\mathcal{G}$. In this case, we say that $\mathbb{P}$ is **faithful** to $\mathcal{G}$.*

Given a graph $\mathcal{G}$, the CMC yields a list of conditional independencies for $\mathbb{P}$ to satisfy. Nevertheless, additional independencies can emerge due to specific parameter choices or constraints, leading to *unfaithful* distributions. A further discussion about these conditions connecting causality to probability can be found in [41]. We conclude by defining *causal sufficiency*, which is a theoretical assumption that is rarely satisfied in real-world settings, despite being often required to ensure CD identifiability.

Definition 3. *A set $\mathbf{X}$ of rvs is **Causally Sufficient (CS)** iff every common cause of any two or more variables in $\mathbf{X}$ is in $\mathbf{X}$.*

2.2 Causal Discovery Benchmark

In this work, we consider 8 classical CD algorithms, which can be classified into 4 categories: constraint-based, score-based, functional-based, and permutation-based. All these methods are implemented in the Python library `causal-learn`, which we will use for the results section. The code and datasets are all publicly available on GitHub. Below, we provide a brief description of each considered CD algorithm, of which a summary is given in Table 1.

Constraint-based. Constraint-based methods rely on statistical tests for conditional independence to discover the causal graph. Among constraint-based methods, we consider the following candidates:

PC. Peter-Clark (PC) [41] is one of the most popular CD algorithms. It starts with a complete undirected graph and recursively prunes off edges based on conditional independence. PC assumes **Faithfulness** and **Causal Sufficiency**. Its output is a Partially Directed Graph, consisting of both directed and undirected edges, undirected edges being those for which PC wasn't able to establish any orientation.

FCI. Fast Causal Inference (FCI) [42] is a relaxation of the PC algorithm that does not assume **Causal Sufficiency**. FCI outputs a Partial Ancestral Graph, which can contain different types of edges, including partially oriented edges or bidirected edges, which indicate the possible existence of unobserved common causes (confounders).

Several conditional independence tests can be used, and should be carefully considered depending on the data. Specifically, the χ^2-test and G^2-test [44] are better suited for discrete data. For continuous data, we should use e.g., the Fisher-z test [15], which is particularly well suited for data with linear causal relationships and a Gaussian distribution, or a kernel-based test [48], which is more flexible in terms of the expected functional form of the causal relationship.

Score-based. Score-based algorithms rely on the definition of a score for measuring the goodness of fit of the graph with respect to the observed distribution. We consider the following score-based methods:

GES. Greedy Equivalent Search (GES) [8] starts from an empty graph, and iteratively adds edges entailing an improved score result (forward phase). It then performs a backward search, where edges are removed whenever this operation implies a better scoring. GES assumes **Faithfulness** and **Causal Sufficiency**.

ExactSearch. ExactSearch [39] relies on decomposable scores and uses dynamic programming to compute *every* local score (i.e., a score for each pair of variables and its candidate set of parents). `causal-learn` also implements a refined version of ExactSearch (A^*) [47], which further improves the search method efficiency.

As for constraint-based algorithms, several scores can be proposed and adapted to the data types. The BIC score [37] is classically proposed, and is particularly well-suited for continuous data with linear causal relationships and Gaussian distributions. For discrete data, the Dirichlet score (BDeu) [5] can be used. Finally, a *general score* [19] based on kernels can be proposed for continuous data, without requiring assumptions on the form of the causal relationships.

Functional-based. Functional-based methods rely on a functional definition of causality, which means specifying a function f which describes the form of the causal dependency: given f, X causes Y is denoted by $Y = f(X) + \epsilon$, where ϵ is a noise variable, independent of X and typically centered. We consider the following methods, based on linear assumptions:

DirectLiNGAM. DirectLiNGAM [38] assumes a *linear non-Gaussian acyclic model* (LiNGAM), meaning a linear function f to describe causality, and a non-Gaussian random noise. DirectLiNGAM performs least squares regressions of each variable on the others, computes the residuals, and tests independence between residuals and variables to finally establish the causal order. It assumes **Causal Sufficiency**.

RCD-LiNGAM. Repetitive Causal Discovery LiNGAM (RCD-LiNGAM) [27] relaxes the causal sufficiency assumptions of DirectLiNGAM, and admits the existence of unobserved variables. As its name suggests, this method is also based on a linear non-Gaussian assumption.

Permutation-based. Within this category, we consider methods based on Sparsest Permutation (SP) [35]. SP proposes to use conditional independence to associate a unique DAG to a causal order of the set of variables, investigated through permutation. The aim is to identify the permutation yielding the sparsest DAG (i.e., the one with the smallest number of edges). SP's identifiability requires **Causal sufficiency** and a strictly weaker **Faithfulness** constraint.

GRaSP. Greedy Relaxations of SP (GRaSP) [23] replaces the exhaustive permutations search of SP with a more efficient permutation space travelling method, making the algorithm more scalable for highly connected graphs.

BOSS. Best Order Score Search (BOSS) [2] introduces and leverages a novel technique known as Grow-Shrink Trees (GST) [2] to construct the unique DAG associated to a permutation, ensuring scalability for high-dimensional data.

Permutation-based methods also require the specification of a score to assess the quality of the considered graphs. The same scores already discussed for score-based methods apply here.

Algorithm	**Type**	**Faithfulness**	**CS**	**Reference**
PC	Constraint-based	✓	✓	[41]
FCI		✓	✗	[42]
GES	Score-based	✓	✓	[8]
ExactSearch		✓	✓	[39] [47]
DirectLiNGAM	Functional-based	✗	✓	[38]
RCD-LiNGAM		✗	✗	[27]
GRaSP	Permutation-based	✗	✓	[23]
BOSS		✗	✓	[2]

Table 1. CD algorithms benchmark. (**CS** = Causal Sufficency)

2.3 Consensus Causal Model

To build our consensus causal model, we first perform a homogenisation step, and map all output causal graphs to a universal causal graph representation characterized through a binary adjacency matrix, enabling coherent information exchange between CD algorithms and aggregation of their heterogeneous outputs. Indeed, as detailed in Sec. 2.2, CD algorithm outputs may present different types of predicted edges (directed, bidirected, undirected, partially directed). For the sake of interoperability, bidirected and undirected edges are mapped into two symmetric 1 entries in the corresponding binary adjacency matrix, under the convention that symmetric entries in the consensus matrix will be interpreted as an identified association between the two involved nodes, which can be either causal (directed or mediated) or anti-causal (e.g., through latent confounding); symmetric entries on the consensus will be graphically represented as undirected. On the other hand, partially directed edges are homogenized to directed,

to strengthen the information contained therein of a forbidden direction. A similar approach was adopted e.g. in [36], for scoring purposes and comparison against ground truth.

Given a set of homogenized binary adjacency matrices sharing the same edge-type conventions, the consensus causal model is computed by their arithmetic average. This leads to a weighted consensus graph, where weights denote edge observation frequencies, and bring information on the consistency of the discovered causal associations. It is worth noticing that our choice of a binary mapping is particularly effective here since it naturally encourages orientation over simple skeleton discovery. Finally, for the sake of a cleaner graphical representation, we consider a majority voting strategy to solve orientation conflicts, whenever possible: we retain the edge direction corresponding to the larger weight between the two possible directions, and keep an edge undirected whenever the two directions have equal weight. Of note, our consensus graph is not constrained to be acyclic, and may allow for cyclic relationships if not actively prohibited (only self-loops are forbidden by construction). Nevertheless, consensus weights provide valuable information for edge pruning. Fig. 1 schematically illustrates how the consensus is built.

Aggregating multiple causal graphs has already been proposed (e.g., [10]), but this procedure is typically performed by adding/removing edges based on a threshold (typically set to 50%). The resulting graph is then interpreted as a binary structure. On the contrary, our consensus model learns from all considered CD methods and combines all discovered edges within a weighted structure that conveys a heuristic measure of confidence: higher (respectively, lower) weights indicate more (respectively, fewer) occurrences across the set of causal graphs. The resulting consensus structure is both denser and more informative, providing potentially new insights together with an associated level of uncertainty. While the main focus of the current study concerns the analysis of the lung cancer data described in the following section, Appendix 5.3 contains some exemplifying results in a controlled simulation study, comparing the performance of the consensus graph with individual algorithms. Overall, experiments conducted on synthetic data show that our consensus graph and its weighted structure effectively allows to drive attention towards the most relevant causal structures, reaching performances comparable to the best-scoring models.

2.4 Lung Cancer Data

Dataset Creation. The data extraction and dataset construction follow the same pipeline used in [34]. We combined patient- and tumor-level data obtained from a lung cancer study [6] available on cBioPortal [16] with genomic information derived from somatic mutations and with the hierarchical organization of cellular pathways provided by Reactome [30]. Using the concept of Super Pathways [13], which groups cellular functions into hierarchical structures, we intersected their gene sets with each patient's mutated driver genes. Only mutations in known driver genes [11,29] were considered, reducing the dimensionality of the genomic feature space and increasing the biological relevance of the associations. For each patient and for each Super Pathway or Sub-pathway, we evaluated whether the intersection between mutated genes and pathway genes was statistically significant, retaining only statistically significant associations. Table 2 shows

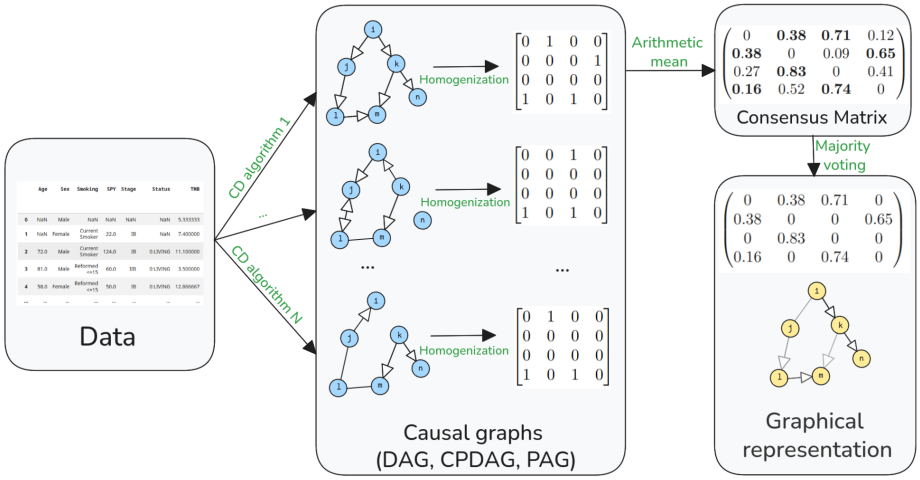


Fig. 1. Diagram explaining the construction of the consensus causal model

the cellular pathways that have been finally retained, and the reference number we will use for graphical representations in Sec. 3. This procedure yielded a reduced and structured dataset that integrates clinical and tumor information (6 features), and genomic data (12 cellular pathways), illustrated in Table 3. We dispose of 1044 subjects in total.

Ref.	Description	Super/ Sub
SP1	Cell-Cell communication	Super
SP2	Cellular responses to stimuli – Oncogene Induced Senescence	Sub
SP3	Chromatin organization – PKMTs methylate histone lysines	Sub
SP4	Developmental Biology	Super
SP5	Extracellular matrix organization	Super
SP6	Gene expression (Transcription) – Transcriptional regulation by RUNX3	Sub
SP7	Immune System – Dectin-2 family	Sub
SP8	Metabolism of proteins – SUMOylation	Sub
SP9	Neuronal System	Super
SP10	Programmed Cell Death – Activation of NOXA and translocation to mitochondria	Sub
SP11	Signal Transduction	Super
SP12	Vesicle-mediated transport – Scavenging by Class H Receptors	Sub

Table 2. Description of each super-/sub- cellular pathway retained for this study.

Further details on the variable reduction, subgroup decomposition, and aggregation steps can be found in [34], which fully documents the original pipeline.

Missing data handling. Most existing CD algorithms need to be fed with complete datasets (*i.e.*, with no missing entries), to learn the underlying causal DAG. While the

	Patient- and tumor-level data						Genomic level data			
Patient	**Sex**	**Age**	**Smoking**	**SPY**	**Stage**	**Status**	**SP1**	**SP2**	$\cdots$	**SP12**
p1	Male	41	Ref$>$15	20	IA	Living	2	0	0	0
p2	Male	51	Ref$\leq$15	10	IIIA	Deceased	0	5	0	1
p3	Female	65	Ref	12	IIIB	Living	2	0	7	0
...	Female	83	Current	32	V	Deceased	1	0	0	1
p1044	Male	67	Ref$>$15	40	IIA	Living	3	2	1	0

Table 3. Dataset encompassing patient- and tumor-level data, adapted from [34]. **SPY**: Smoking Pack Years. **SP**: Super-/Sub-pathway. *Sex*, *Smoking*, *Status*, and *Stage* are by nature binary or categorical variables, while the remaining variables are continuous (Ref = Reformed).

cellular pathways data is complete (a missing record implies no mutation has been observed for this sample, so we can safely replace it with 0), the 6 features reporting patient- and tumor-level data are of heterogeneous nature (continuous and categorical) and contain up to 20% of missing values (see Supplementary Fig. 6). A simple, straightforward solution consists of removing every sample containing at least one missing entry. Nevertheless, this entails a significant loss of information, since more than 30% of the samples are partially corrupted. An alternative approach implies performing missing data imputation prior to the CD task. Several methods of missing data imputation exist, and can be broadly classified into two categories: univariate and multivariate imputation. The first one independently considers one feature at a time and uses the observed values in each feature containing missing entries to replace them, such as the *mean imputation*. Multivariate imputation instead uses all the observed features in the dataset to impute missing values, such as the k-Nearest Neighbors (k-NN) algorithm [3], which we decided to use in this work. To account for heterogeneous data types, we chose the Heterogeneous Euclidean-Overlap Metric (HEOM) [45] to define the nearest neighbors.

Definition 4. *Let $\mathbf{x} = (x_d)_{d=0,\dots,D}$ and $\mathbf{y} = (y_d)_{d=0,\dots,D}$ be two D-dimensional samples, composed of continuous and categorical features. Their HEOM distance, $d_{\mathrm{HEOM}}(\mathbf{x}, \mathbf{y})$, is defined as follows:*

$$d_{\mathrm{HEOM}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{d=1}^{D} d(x_d, y_d)^2}, \text{where } d(x_d, y_d) = \begin{cases} \begin{cases} 0 \text{ if } x_d = y_d \\ 1 \text{ otherwise} \end{cases} & \text{if } d \text{ is categorical} \\ \dfrac{|x_d - y_d|}{\max_d - \min_d} & \text{if } d \text{ is continuous} \end{cases}$$

$\max_d$ (resp. $\min_d$) are the maximum (resp. the minimum) of the d-th feature across all samples.

We equipped the k-NN algorithm with the HEOM distance to find the k nearest neighbors ($k = 5$ in this work) of each sample $\mathbf{x} = (x_d)_{d=0,\dots,D}$ with a missing entry for at least one feature, x_{d^*} in our dataset. If x_{d^*} is continuous, we impute its value using the mean of the $d^{*\mathrm{th}}$ feature over the set of nearest neighbors; otherwise, we impute by the most frequent value. This process has the advantage of preserving the original distribution of the data (see Supplementary Fig. 5).

In Sec. 3 we will refer to the dataset from which partially observed samples have been removed as *Filtered*, and the dataset in which missing entries have been imputed as *Imputed*.

Prior knowledge. Some CD algorithms allow for the injection of prior knowledge about existing (resp. forbidden) causal relationships: this can greatly facilitate the causal discovery task in terms of both accuracy and computational load. Concerning our specific application, we dispose of limited, yet informative, prior information concerning some variables about patient and tumor data. Specifically, we can assert that *Age* and *Sex* should have no parents variable (*i.e.*, their in-degree is 0), while *Status* should have no children (*i.e.*, its out-degree is 0) as the patients' outcome should not cause any of the variables considered. In Sec. 3 we will associate the mention *prior* whenever this prior knowledge has been provided as additional input to the CD algorithm.

3 Results

3.1 Individual causal models display substantial heterogeneity

We generated a total of 44 causal graphs relating the 6 features of patient and tumor data and the 12 cellular pathways. Each graph resulted from the application of a CD algorithm with specific hyperparameter choices (i.e., the conditional independent test or the score, as described in Sec. 2.2). To prevent over-weighting a given CD method, we deduplicate identical causal graphs whenever they result from the same method run with different hyperparameter choices, leading to a final population of 38 candidate individual causal models. We propose several strategies to merge the obtained causal graphs, detailed in Tab. 4. We build a consensus causal model for each strategy.

Strategy	**Which graphs to be merged together?**	$\mathbf{n}$
St.1	All causal graphs	38
St.2	Causal graphs obtained using prior knowledge	9
St.3	Causal graphs obtained on the *Filtered* dataset	18
St.4	Causal graphs obtained on the *Imputed* dataset	18

Table 4. Strategies for causal graph merging. $\mathbf{n}$ denotes the number of graphs included in each set.

Fig. 2 shows several interesting metrics evaluated within each set of causal graphs from Tab. 4, which offer a quantitative measure of the observed heterogeneity across individual causal models. We firstly consider the Structural Hamming Distance (SHD, top left panel), a reference distance to compare the structure of two (partially) directed graphs, which counts the number of edge deletions, insertions, or flips needed to turn one graph into the other. We evaluate the SHD between each pair of causal graphs within every merging strategy. One can see that the minimum obtained SHD between any 2 discovered causal graphs is 25 for *St.2*, with an average SHD above 30 for the other merging strategies. This highlights the high variability of the discovered causal

structures and, consequently, emphasizes the challenge of relying on a specific causal graph candidate. Supplementary Fig. 7 further supports the heterogeneity statement.

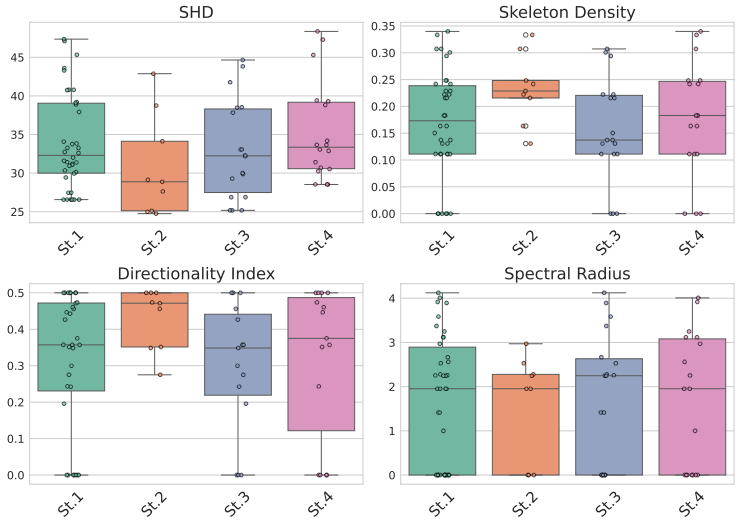


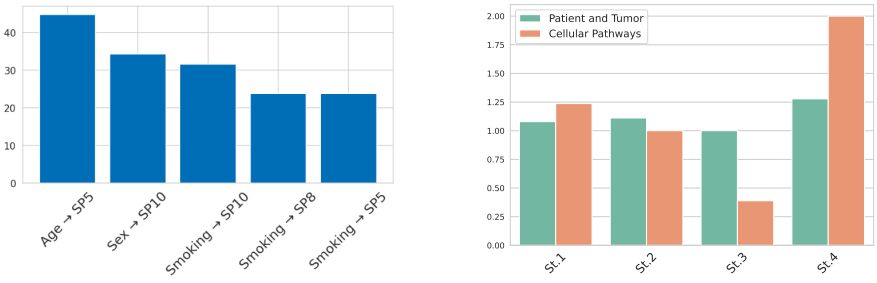
Fig. 2. Boxplot reporting some graph metrics per merging strategy. (Top left) Structural Hamming distance (SHD). (Top right) Skeleton density. (Bottom left) Directionality index. (Bottom right) Spectral radius.

Fig. 2 also conveys another relevant piece of information: the importance of prior knowledge. Indeed, causal graphs in *St.2*, obtained by specifying the role of some strategic nodes (Sec. 2.4), display a significantly higher directionality index (the ratio between purely directed *vs* undirected edges) and skeleton density (the overall number of discovered edges) with respect to the other strategies. This suggests that the injected prior knowledge has strongly contributed to orienting edges, while supporting CD in robustly identifying meaningful relationships across the observed variables. To further stress the relevance of prior knowledge, we have also assessed the spectral radius of all discovered causal graphs. *St.2* shows a milder spectral radius: of note, a small spectral radius corresponds to a graph closer to being acyclic, and a small total vertex degree.

Finally, one can observe that the imputation process performed in *St.4* does not seem to lead to more stable discovered causal structures compared to *St.3*, where incomplete entries have been removed, with a consequent significant reduction of the dataset size. One possible explication can come from the nature of the missing mechanism, whose analysis is out of the scope of the current work.

3.2 The consensus causal model offers rich biological insights

We build the consensus causal model following the procedure described in Sec. 2.3. We use St. 2 to build the consensus, because, as commented in Sec. 3.1, it shows a



(a) (b)

Fig. 3. (a) Most frequent edges (%) connecting patient and tumor level data with cellular pathways variables in *St.1*. (b) Average out-degree of the 'Smoking' variable by strategy stratified by patient and tumor variables vs cellular pathways variables.

significantly higher directionality index. In Fig. 4, we compare the consensus model, where edge weights are denoted by color intensity, with the causal graph obtained by [34] applying PC. For the sake of completeness, Supp. Fig. 7 further shows the full weighted consensus matrices as heatmaps, before applying the voting strategy for orientation. The consensus causal model appears substantially denser than the reference one, in some cases explaining some previously found direct associations (e.g., *SP12*—*SP4*) through more complex paths. This opens up a deeper discussion on the complex network relating cellular pathways and patient-level information in lung cancer.

Several edges observed in the consensus model are supported by previous external studies. For instance, in clinical oncology, cancer stage is one of the strongest predictors of survival, with advanced stages generally associated with poorer outcomes. In the consensus causal model, we observe a collider structure *Stage* $\rightarrow$ *Status* $\leftarrow$ *Smoking*, meaning the edge *Smoking* $\rightarrow$ *Status* persists even after conditioning on stage and all other available features. This pattern suggests that smoking contributes independently to survival outcomes rather than acting solely through the stage progression. Evidence from independent studies supports this interpretation, since continued smoking after cancer diagnosis has been associated with significantly higher mortality [7,9]. This causal association was not identified in the individual causal model from [34].

The causal relationships from the patient and tumor level data towards cellular pathways are mostly mediated by *Age*, *Sex*, and *Smoking*, as underlined in Fig. 3 (a). Specifically, more than 40% of the considered individual causal models have identified the causal association *Age*$\rightarrow$*SP5: Extracellular Matrix Organization*, which were not observed in [34]. Extracellular Matrix (ECM) is a large network of proteins and other molecules that surround, support, and give structure to cells and tissues in the body. Previous studies show that aging is strongly linked to changes in the ECM, including reduced remodeling capacity, altered collagen structure, and increased stiffness of the surrounding tissue environment [21,43]. With increasing age, the ECM undergoes irregular and progressive modifications across different tissues, and these alterations are increasingly recognized as contributors to disease states, including cancer [28]. Importantly, the remodeled matrix in aged tissue can create physical and biochemical condi-

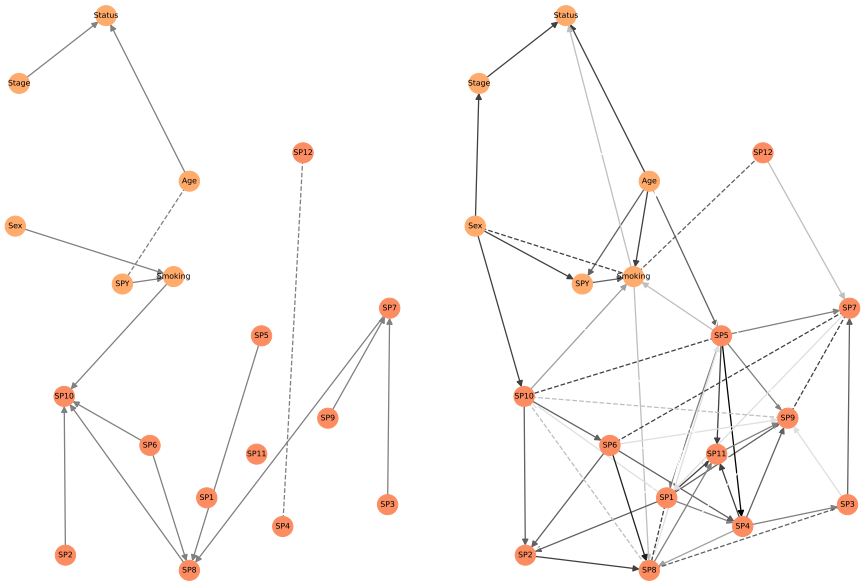


Fig. 4. (Left) Causal graph from [34] and (Right) our consensus causal graph from *St.2*. Dashed edges are undirected edges; edge color intensity in the consensus causal model corresponds to the frequency of the edge occurrence.

tions that support tumor growth and metastatic spread. In addition, these ECM changes may act as a barrier to therapeutic response, limiting the effectiveness of both conventional treatments and immune-based interventions [4]. Similarly, the causal association *Sex→SP10: Programmed Cell Death – Activation of NOXA and translocation to mitochondria*, absent in the reference graph from [34], has been recovered by over 30% of individual causal discovery algorithms queried for consensus. The existence of this causal relationship is supported by several external works (e.g., [25]), highlighting sex differences in cell death pathways, including in cancer, with implications on sex-related differences in anti-programmed cell death protein 1 therapy [12].

Also, smoking habits are identified as strongly altering several cellular pathways, underlying the key role it plays in lung cancer: this central role is further stressed by Fig. 3 (b), where we report the average out-degree of *Smoking* per merging strategy. Among consistently discovered causal relationships, *Smoking→SP10* was already observed in [34]. It is well documented that cigarette smoke causes mitochondrial damage and can trigger intrinsic cell-death programs [14,40]. Meanwhile, NOXA is a stress-responsive pro-apoptotic protein that acts at the mitochondria. When activated, it helps initiate mitochondrial cell death by blocking survival proteins, which ultimately triggers apoptosis [31,1]. Our consensus model also suggests that *Smoking* causally influences

the cellular pathway *SP8: Metabolism of proteins – SUMOylation*. SUMOylation is a general protein-modification process that helps regulate fundamental cellular functions such as cell growth, migration, responses to stress, and processes related to pulmonary diseases, including cancer [46,49]. Studies have shown that exposure to cigarette smoke alters SUMOylation patterns in different cell types, including lung cells and immune cells, indicating that smoking can directly interfere with this protein-regulation system [18,50]. This relationship was missing in the individual model from [34]. Of note, in the consensus model plot in Fig. 4, the reverse direction (*SP10→Smoking*) is reported due to the voting strategy used for edge orientation. Nevertheless, as depicted in Fig. 3 (a), the more reasonable causal flow *Smoking→SP10* is correctly recovered by more than 30% of the considered CD algorithms: in this case, prior knowledge should have been preferred over voting to discriminate the best orientation. The voting strategy could also have been responsible for some observed edge reversals between our consensus graph and that of [34]. This is, for instance, the case for SP2 and SP10, for which our consensus associates a stronger weight to the *SP10→SP2* direction (0.7) with respect to the other one (0.4).

A remark should be spent on the role within the consensus model (as well as the reference graph by [34]) of the tumor-level nodes *Stage* and *Status*. Surprisingly, there is no discovered direct or indirect relationship between such nodes and any of the cellular pathways, indicating that corruption of the cellular functions is not identified as causally affecting cancer development or the outcome of the patient. This behavior has also been observed when summarizing the information on genetic disregulation into a single node, Tumor Mutational Burden (TMB - see Supplementary Sec. 5.2): TMB is identified in the consensus model as a direct effect of *Smoking*, *Age*, and *Sex*, with no children. This may be partially explained by the fact that our dataset only comprises diseased samples: the inclusion of healthy samples in the study would have possibly allowed highlighting the counterfactuals and hence, other mechanistic relationships between cellular pathways deregulation and lung cancer.

Finally, as noted earlier, the consensus causal graph appears substantially denser than the graph reported in [34], especially in the cellular pathways subgraph. All causal relationships identified between pathways can be explained thanks to the intersection of cancer driver genes: cellular pathways sharing deregulated molecular components are more likely to appear as connected in the causal graph. One example of these patterns is the causal edge observed from *SP10: Programmed Cell Death – Activation of NOXA and translocation to mitochondria* to *SP2: Cellular responses to stimuli – Oncogene Induced Senescence*. In the Reactome hierarchy, these belong to broader functional categories (*Programmed Cell Death* and *Cellular Responses to Stimuli*), while the ones analyzed here represent more specific sub-pathways nested within them. These two sub-pathways share the cancer drivers *TP53* and *TFDP1* [11], which have been associated with regulatory roles in both apoptotic and senescence-related mechanisms. The presence of shared driver genes suggests potential biological cross-talk, and the fact that this relationship appears as a directed edge in our model indicates that the CD algorithms successfully captured this underlying functional dependency. Moreover, all causal pathways across cellular pathways include at least one among *SP5*, *SP8*, *SP10*, *SP12*, all related to *Smoking*.

More generally, despite density might be perceived as a drawback in CD in relation to a higher false positive rate, the trade-off between sensitivity and specificity in causal discovery for biomedical data remains an open question, and largely depends on the intended downstream task. The weighted causal structure of our consensus causal model conveys richer information than a classical binary adjacency matrices, enabling tuning the level of sparsity. For instance, if the graph is intended to be used to guide experimental validation, sparser structures are preferable in order to focus resources on a small number of targets. In this case, only edges with high weights would be retained. Conversely, when aiming for a more comprehensive, system-level understanding, overly aggressive pruning may discard important regulatory relationships: edges with lower weights may still be relevant and should be kept. A related discussion on the trade-off between false positives and false negatives in a medical context is provided in [26], where it motivates the development of a variant of the PC algorithm that allows the user to control the false discovery rate rather than the independence test power. Notably, in our experiments on simulated data (see supplementary Fig. 14), a threshold of 50% provides a reasonable balance between these two objectives.

4 Conclusions

Our work tackles the challenging problem of causal discovery in complex biomedical contexts, where the ground truth is missing, and the compliance of data with the strong assumptions required for causal identifiability are hard to prove. To overcome the difficulties encountered by causal discovery in such contexts, we propose a consensus causal model obtained by aggregating the outputs of multiple causal discovery algorithms based on the consistency of their causal associations. The construction of the consensus departs from the usual binary representation in causal discovery in favor of a continuous perspective, which enables a heuristic quantification of uncertainty in the inferred relationships. The case study of lung cancer allows us to demonstrate the effectiveness of our ensemble strategy as the consensus model recovers most causal relationships previously established by [34] while also providing additional rich and biologically valid insights into the causal relationships linking *Age* and *Smoking* to cellular functions. Overall, the proposed consensus representation aggregates multiple levels of causal evidence into a single object, thereby offering greater flexibility and interpretability for downstream applications. Despite the consensus model already yields an informal measure of confidence, in future work we aim to fully embrace the paradigm of ensemble causal learning and investigate more refined strategies for merging causal graphs, together with a rigorous approach to uncertainty quantification for causal discovery and mechanisms to enforce acyclicity in the aggregated structure.

References

1. Albert, M.C., Brinkmann, K., Kashkar, H.: Noxa and cancer therapy: Tuning up the mitochondrial death machinery in response to chemotherapy. Molecular & cellular oncology **1**(1), e29906 (2014)

2. Andrews, B., Ramsey, J., Sánchez-Romero, R., Camchong, J., Kummerfeld, E.: Fast scalable and accurate discovery of dags using the best order score search and grow-shrink trees. Advances in Neural Information Processing Systems (2023)
3. Batista, G.E.d.A.P.A., Monard, M.C.: A study of k-nearest neighbour as an imputation method. Soft Computing System: design, management and applications (2002)
4. Boaru, D.L., De Leon-Oliva, D., De Castro-Martinez, P., Garcia-Montero, C., Fraile-Martinez, O., García-González, B., Pérez-González, I., Alhaddadin, M.N.M., Barrena-Blázquez, S., Lopez-Gonzalez, L., et al.: Extracellular matrix dysregulation in aging, calcification, and cancer diseases: insights into cellular senescence, inflammation, and novel therapeutic strategies. International Journal of Biological Sciences **21**(15), 6808 (2025)
5. Buntine, W.: Theory refinement on bayesian networks. In: Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence (UAI 1991). pp. 52–60. Morgan Kaufmann (1991)
6. Campbell, J.D., Alexandrov, A., Kim, J., Wala, J., Berger, A.H., Pedamallu, C.S., Shukla, S.A., Guo, G., Brooks, A.N., Murray, B.A., et al.: Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nature genetics **48**(6), 607–616 (2016)
7. Chellappan, S.: Smoking cessation after cancer diagnosis and enhanced therapy response: mechanisms and significance. Current Oncology **29**(12), 9956–9969 (2022)
8. Chickering, D.M.: Optimal structure identification with greedy search. Journal of Machine Learning (2002)
9. Cinciripini, P.M., Kypriotakis, G., Blalock, J.A., Karam-Hage, M., Beneventi, D.M., Robinson, J.D., Minnix, J.A., Warren, G.W.: Survival outcomes of an early intervention smoking cessation treatment after a cancer diagnosis. JAMA oncology **10**(12), 1689–1696 (2024)
10. Constantinou, A., Kitson, N.K., Liu, Y., Chobtham, K., Amirkhizi, A.H., Nanavati, P.A., Mbuvha, R., Petrungaro, B.: Open problems in causal structure learning: A case study of covid-19 in the uk. Expert Systems with Applications **234**, 121069 (2023)
11. Dressler, L., Bortolomeazzi, M., Keddar, M.R., Misetic, H., Sartini, G., Acha-Sagredo, A., Montorsi, L., Wijewardhane, N., Repana, D., Nulsen, J., et al.: Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the network of cancer genes (ncg) resource. Genome biology **23**(1), 35 (2022)
12. Duma, N., Abdel-Ghani, A., Yadav, S., Hoversten, K.P., Reed, C.T., Sitek, A.N., Enninga, E.A.L., Paludo, J., Aguilera, J.V., Leventakos, K., et al.: Sex differences in tolerability to anti-programmed cell death protein 1 therapy in patients with metastatic melanoma and non-small cell lung cancer: are we all equal? The oncologist **24**(11), e1148–e1155 (2019)
13. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al.: The reactome pathway knowledgebase. Nucleic acids research **46**(D1), D649–D655 (2018)
14. Fetterman, J.L., Sammy, M.J., Ballinger, S.W.: Mitochondrial toxicity of tobacco smoke and air pollution. Toxicology **391**, 18–33 (2017)
15. Fisher, R.A.: On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. Metron **1**, 3–32 (1921)
16. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al.: Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. Science signaling **6**(269), pl1–pl1 (2013)
17. Glymour, C., Zhang, K., Peter, S.: Review of causal discovery methods based on graphical models. frontiers in Genetic (2019)
18. Gross, T.J., Powers, L.S., Boudreau, R.L., Brink, B., Reisetter, A., Goel, K., Gerke, A.K., Hassan, I.H., Monick, M.M.: A microrna processing defect in smokers' macrophages is linked to sumoylation of the endonuclease dicer. Journal of Biological Chemistry **289**(18), 12823–12834 (2014)

19. Huang, B., Zhang, K., Lin, Y., Schölkopf, B., Glymour, C.: Generalized score functions for causal discovery. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1551–1560 (July 2018)
20. Joshi, S., Urteaga, I., Van Amsterdam, W.A., Hripcsak, G., Elias, P., Recht, B., Elhadad, N., Fackler, J., Sendak, M.P., Wiens, J., et al.: Ai as an intervention: improving clinical outcomes relies on a causal approach to ai development and validation. Journal of the American Medical Informatics Association **32**(3), 589–594 (2025)
21. Kehlet, S.N., Willumsen, N., Armbrecht, G., Dietzel, R., Brix, S., Henriksen, K., Karsdal, M.A.: Age-related collagen turnover of the interstitial matrix and basement membrane: Implications of age-and sex-dependent remodeling of the extracellular matrix. PloS one **13**(3), e0194458 (2018)
22. Kotoku, J., Oyama, A., Kitazumi, K., Toki, H., Haga, A., Yamamoto, R., Shinzawa, M., Yamakawa, M., Fukui, S., Yamamoto, K., et al.: Causal relations of health indices inferred statistically using the directlingam algorithm from big data of osaka prefecture health checkups. Plos one **15**(12), e0243229 (2020)
23. Lam, W.Y., Andrews, B., Ramsey, J.: Greedy relaxations of the sparsest permutation algorithm. In The 38th Conference on Uncertainty in Artificial Intelligence (2022)
24. Le, T.D., Liu, L., Tsykin, A., Goodall, G.J., Liu, B., Sun, B.Y., Li, J.: Inferring microrna–mrna causal regulatory relationships from expression data. Bioinformatics **29**(6), 765–771 (2013)
25. Li, H., Pin, S., Zeng, Z., Wang, M.M., Andreasson, K.A., McCullough, L.D.: Sex differences in cell death. Annals of neurology **58**(2), 317–321 (2005)
26. Li, J., Wang, Z.J.: Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. Journal of Machine Learning Research **10**, 475–514 (2009)
27. Maeda, T.N., Shimizu, S.: Rcd: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In International Conference on Artificial Intelligence and Statistics (2020)
28. Marino, G.E., Weeraratna, A.T.: A glitch in the matrix: Age-dependent changes in the extracellular matrix facilitate common sites of metastasis. Aging and cancer **1**(1-4), 19–29 (2020)
29. Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., et al.: A compendium of mutational cancer driver genes. Nature Reviews Cancer **20**(10), 555–572 (2020)
30. Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B., et al.: The reactome pathway knowledgebase 2024. Nucleic acids research **52**(D1), D672–D678 (2024)
31. Morsi, R.Z., Hage-Sleiman, R., Kobeissy, H., Dbaibo, G.: Noxa: role in cancer pathogenesis and treatment. Current cancer drug targets **18**(10), 914–928 (2018)
32. Pearl, J.: Causality. Cambridge university press (2009)
33. Prosperi, M., et al.: Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nature Machine Intelligence **2**(7), 369–375 (2020)
34. Ramos, R.H., Simao, A., Mousavi, M.R.: Causal model discovery in cancer guided by cellular pathways. In: International Conference on Computational Methods in Systems Biology. pp. 174–195. Springer (2024)
35. Raskutti, G., Uhler, C.: Learning directed acyclic graphs based on sparsest permutations. Stat (2018)
36. Rehak, J., Falkenstein, A., Doehner, F., Beyerer, J.: Metrics for the evaluation of learned causal graphs based on ground truth. In: ML4CPS–Machine Learning for Cyber-Physical Systems. UB HSU (2024)
37. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2), 461–464 (1978)

38. Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., Hoyer, P.O., Bollen, K.: Directlingam: A direct method for learning a linear non-gaussian structural equation model. Journal of machine learning (2011)
39. Silander, T., Myllymaki, P.: A simple approach for finding the globally optimal bayesian network structure. In Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (2006)
40. Song, Q., Chen, P., Liu, X.M.: The role of cigarette smoke-induced pulmonary vascular endothelial cell apoptosis in copd. Respiratory Research **22**(1), 39 (2021)
41. Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, prediction, and search. MIT press (2000)
42. Spirtes, P., Meek, C., Richardson, T.: Causal inference in the presence of latent variables and selection bias. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, Montreal, Quebec, Canada, August 18–20, 1995 (pp. 499–506). Morgan Kaufmann. (1995)
43. Statzer, C., Park, J.Y.C., Ewald, C.Y.: Extracellular matrix dynamics as an emerging yet understudied hallmark of aging and longevity. Aging and Disease **14**(3), 670 (2023)
44. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing bayesian network structure learning algorithm. Machine Learning **65**(1), 31–78 (2006)
45. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. Journal of artificial intelligence research **6**, 1–34 (1997)
46. Yang, Y., He, Y., Wang, X., He, G., Zhang, P., Zhu, H., Xu, N., Liang, S., et al.: Protein sumoylation modification and its associations with disease. Open biology **7**(10) (2017)
47. Yuan, C., Malone, B.: Learning optimal bayesian networks:a shortest path perspective. Journal of Artificial Intelligence Research (2013)
48. Zhang, K., Peters, J., Janzing, D., Schölkopf, B.: Kernel-based conditional independence test and application in causal discovery. In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011). pp. 804–813. AUAI Press (July 2011)
49. Zheng, X., Wang, L., Zhang, Z., Tang, H.: The emerging roles of sumoylation in pulmonary diseases. Molecular Medicine **29**(1), 119 (2023)
50. Zhou, H., Zhang, L., Li, Y., Wu, G., Zhu, H., Zhang, H., Su, J.K., Guo, L., Zhou, Q., Xiong, F., et al.: Cigarette smoke extract stimulates bronchial epithelial cells to undergo a sumoylation turnover. BMC Pulmonary Medicine **20**(1), 276 (2020)

5 Appendix

5.1 Supplementary Figures

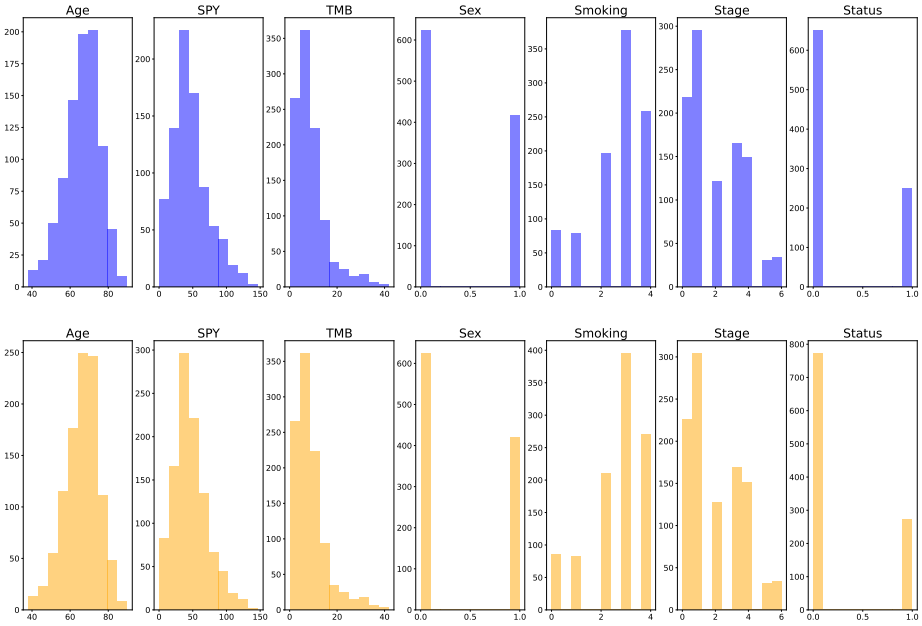


Fig. 5. Data distribution. Top line: original dataset. Bottom line: distribution after imputation.

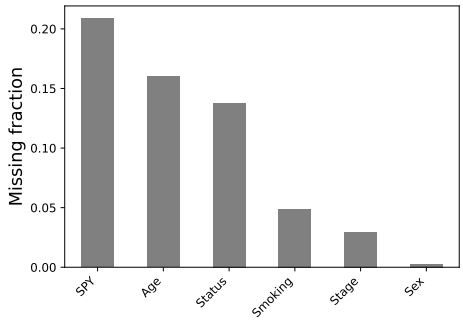


Fig. 6. Prevalence of missing data across all features within the patient and tumor dataset.

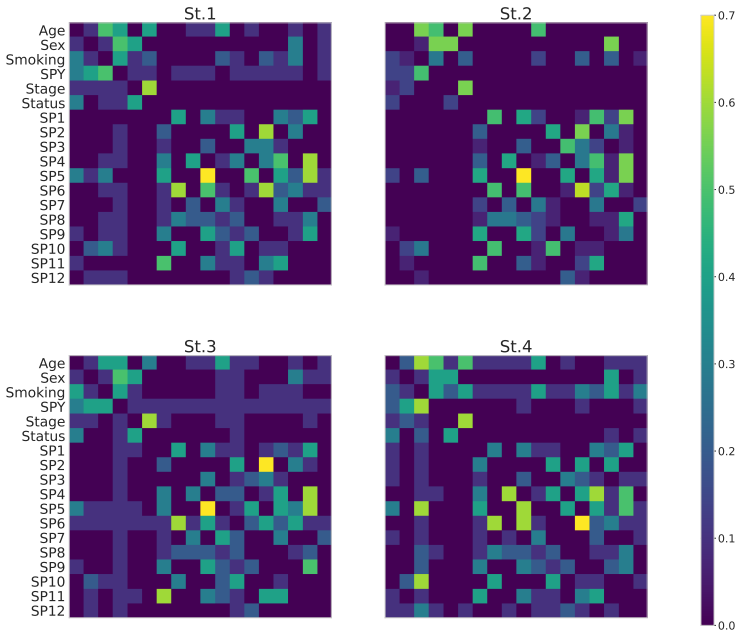


Fig. 7. Mean adjacency matrices per merging strategy on the full dataset (patient and tumor data, and genetic level information). Causality direction corresponds to: row $\rightarrow$ column. A brighter color denotes a higher occurrence of the corresponding edge in the considered set of causal graphs. Two strongly connected components emerge, one relating patient and tumor level data and one relating cellular pathways.

5.2 A focus on patient- and tumor- level variables

Here, we restrict ourselves to the 6 features of patient and tumor level data, and additionally consider Tumor Mutational Burden (TMB) as a proxy of the total degree of genetic mutations affecting cellular pathways, available for all samples in our dataset. As for the main study, consensus causal models are built following the merging strategies introduced in Sec. 3.

Fig. 8 shows the mean adjacency matrices within each group of causal graphs, highlighting again the variability of the discovered causal structures. This can be further appreciated by looking at Fig. 9, upper left panel, which shows pairwise Structural Hamming Distance (SHD). As for the main study, a straightforward conclusion one can draw from the results shown in Fig. 9 is the importance of injecting some (limited) prior knowledge into the CD approach. Indeed, causal graphs analysed in *St.2* (with priors) display a significantly high skeleton density and directionality index, coupled with a smaller spectral radius, suggesting the discovery of a more stable acyclic and still dense graph. Fig. 10 also shows the fraction of strongly connected graphs (i.e., where each node is accessible from any other node in the graph): none of the graphs discovered using prior knowledge are strongly connected, while approximately 5.5% of the graphs from the other merging strategies are.

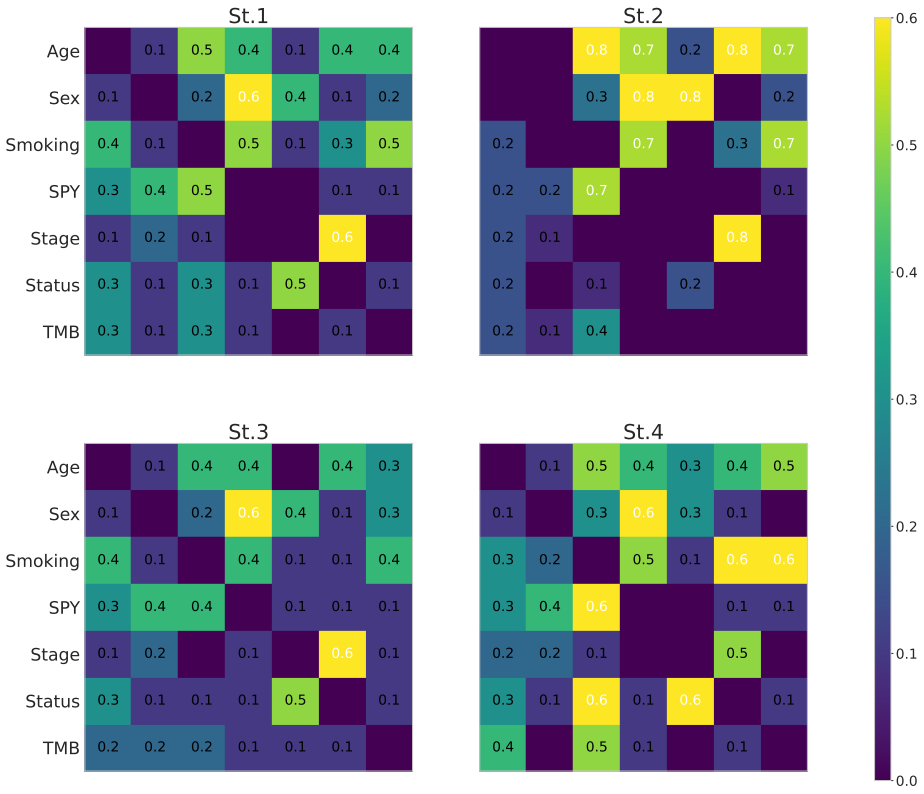


Fig. 8. Mean adjacency matrices per merging strategy on the full dataset (patient and tumor data, and genetic level information). Causality direction corresponds to: row $\rightarrow$ column. A brighter color denotes a higher occurrence of the corresponding edge in the considered set of causal graphs.

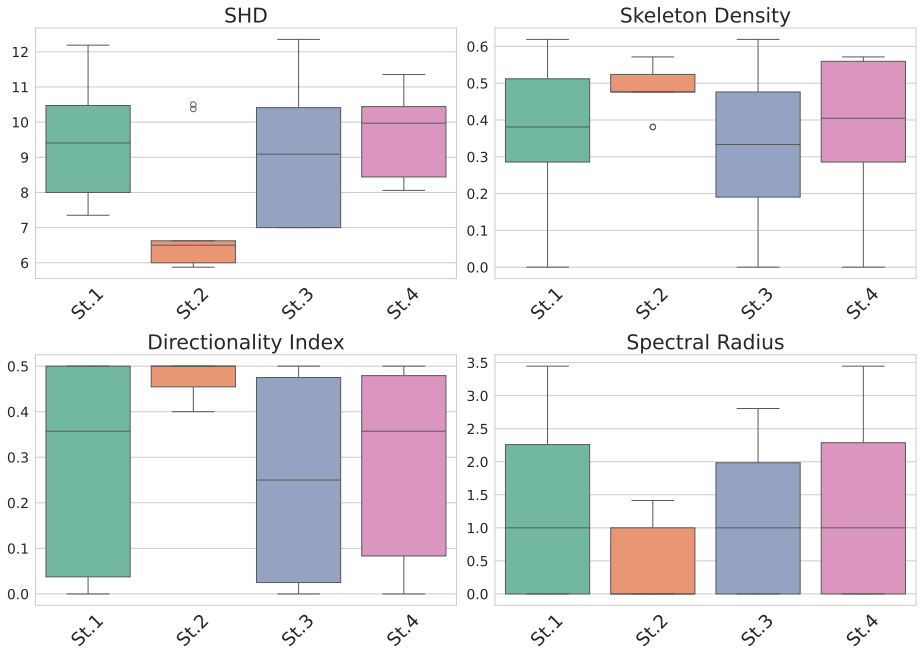


Fig. 9. Boxplot reporting some graph metrics per merging strategy. (Top left) Structural Hamming distance (SHD). (Top right) Skeleton density. (Bottom left) Directionality index. (Bottom right) Spectral radius.

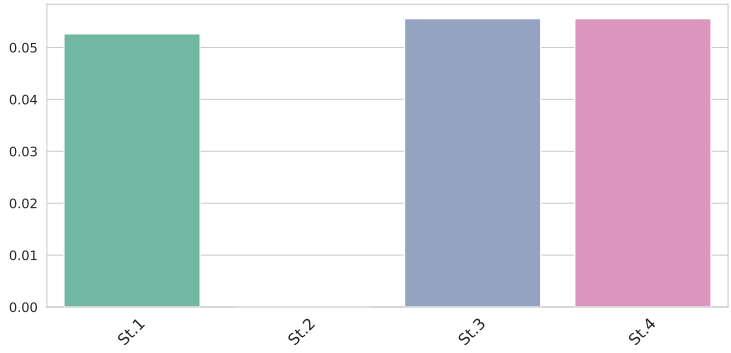


Fig. 10. Fraction of strongly connected graphs per merging strategy

Finally, Fig. 11 compares the causal graph obtained in the original work by [34] using PC with our consensus causal model obtained using *St.2*. Again, the reference graph appears as a subgraph of our consensus model, and misses some very relevant causal associations, such as the collider structure *Stage* $\rightarrow$ *Status* $\leftarrow$ *Smoking*, already commented in Sec. 3.2. Moreover, one can clearly see that *Smoking*, *Age*, and *Sex* are all parent nodes of *TMB*, thus supporting their role in the causal deregulation of cellular pathways.

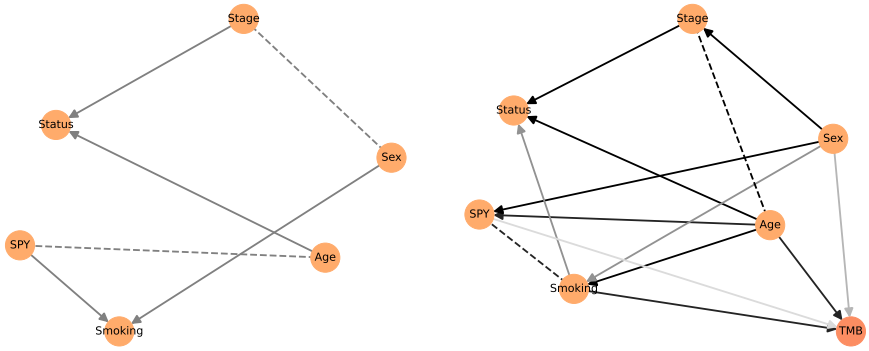


Fig. 11. (Left) Causal graph from [34] and (Right) our consensus causal graph from *St.2*. Dashed edges denote undirected edges; edge color intensity in the consensus causal model corresponds to the frequency of the edge occurrence. [34] did not consider TMB in their study, so it only appears on the consensus model.

5.3 Controlled simulation experiments

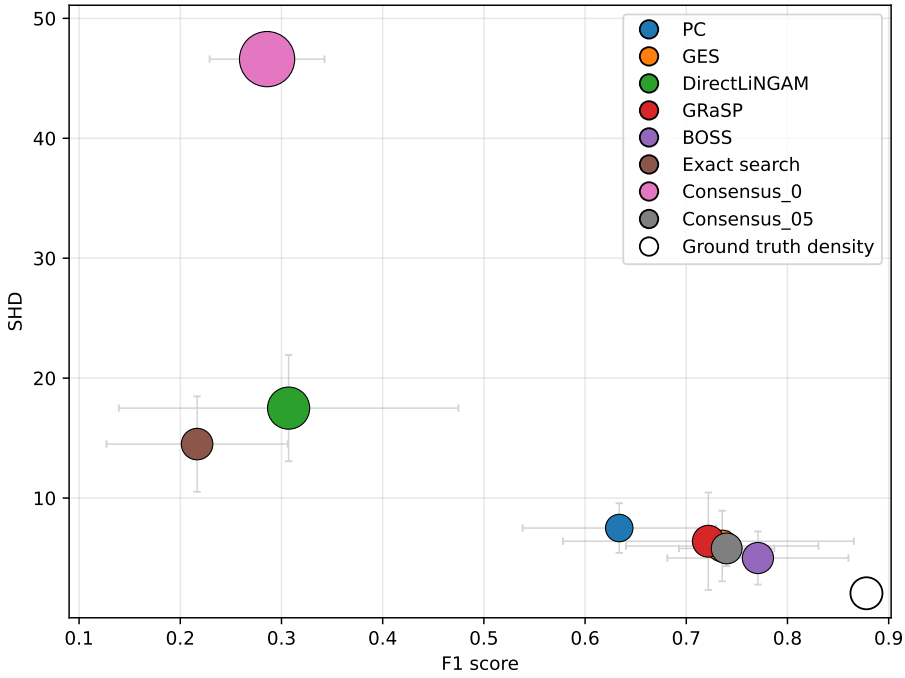


Fig. 12. SHD and F1 score comparing the benchmark CD algorithms and the consensus one. 10 independent datasets of 1000 samples each were randomly generated from 8-node causal graphs, using linear causal relationship ($f(x) = x$), and independent centered Gaussian noise with standard deviation $\sigma = 0.1$. Filled dots represent the mean score over the 10 runs for each method, and error bars represent the corresponding standard deviations. The size of each filled dot is proportional to the average density of the algorithm's output. The empty dot in the top-right corner provides a measure of the average density of the ground truth, as a reference.

To assess the performance of our consensus approach, we perform controlled experiments using synthetic data, with a known ground-truth causal graph, since in real-world settings, this is usually not available. To do so, we proceed as follows :

1. We generate a DAG by randomly generating its adjacency matrix using Erdős–Rényi model for DAGs, and ensuring acyclicity. To do so, we specify the mean density of the desired causal graph through a parameter $p \in]0,\ 1[$, that corresponds to the parameter of a Bernoulli law for each entry in the adjacency matrix.
2. Given the DAG, samples for each variable X_i are obtained from its parents' set PA_i through a functional transformation f_i and a random independent additive

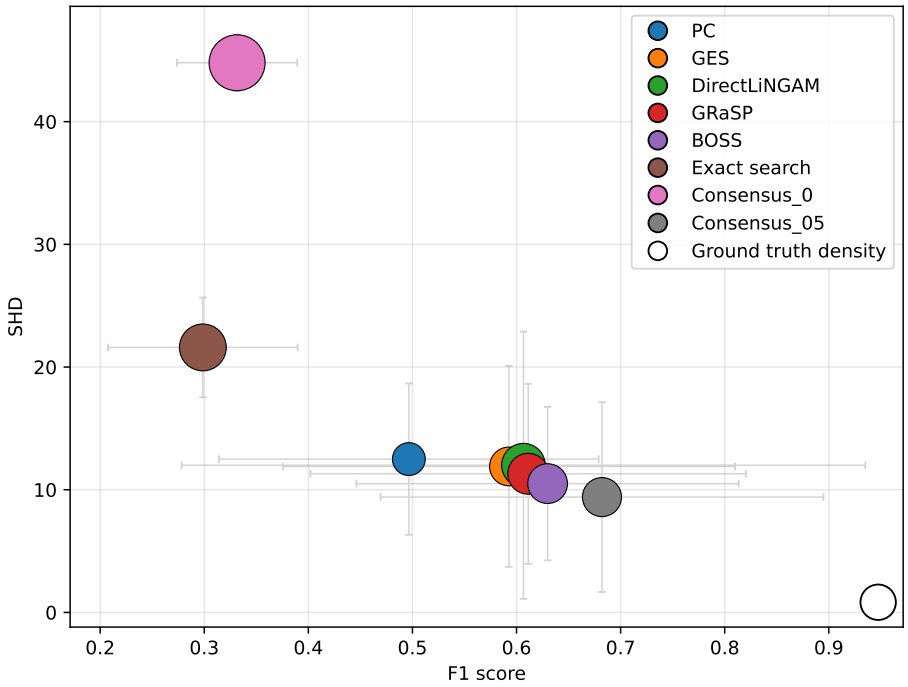


Fig. 13. SHD and F1 score comparing the benchmark CD algorithms and the consensus one. 10 independent datasets of 1000 samples each were randomly generated from 8-node causal graphs, using non-linear causal relationship (f is randomly chosen among non-linear functions), and independent Uniform noise scaled by $\sigma = 0.1$. Filled dots represent the mean score over the 10 runs for each method, and error bars represent the corresponding standard deviations. The size of each filled dot is proportional to the average density of the algorithm's output. The empty dot in the bottom-right corner provides a measure of the average density of the ground truth, as a reference.

noise variable ϵ_i:

$$X_i = \sum_{X_j \in PA_i} f_i(X_j) + \epsilon_i \tag{2}$$

In our experiments, we use $f_i = f$, for different specifications of f.

In Figures 12-13 we show results from two simulation studies. In both cases, we fixed the number of variables in each random graph to 8, and the parameter of the Bernoulli distribution to $p = 0.4$, which means that the average number of edges in the ground truth is $\frac{8(8-1)}{2}0.4 = 11.2$. Figure 12 refers to the simple case of linear Gaussian data, while Figure 13 refers to the non-linear non-Gaussian one, in the case of the non-linear non-Gaussian, the function f in (2) is randomly chosen among the following functions: ReLU, sigmoid, logarithmic, exponential, sinus, hyperbolic tangent and quadratic. For each setting, we randomly generated 10 datasets following the steps

described above, and compare each individual causal discovery algorithm and our consensus model against the ground truth by reporting the Structural Hamming Distance (SHD), the F1 score, and a relative measure of the density of the output graph. Since all features were continuous in all synthetic dataset discussed here, PC were run with the Fisher-Z independence test, while the BIC score was used for GES, GRaSP, and BOSS. Finally, we used the search method 'astar' for Exact search.

As a reference, in both Figures, we plot the score of the consensus model where each edge was kept irrespective of its weights (Consensus_0, in pink), which, as expected, obtains a high SHD and low F1 score due to its high density. Nevertheless, one can clearly see that if we fixed a 0.5 threshold for edge pruning, the consensus model (Consensus_05, in gray) significantly improves both scores, indicating that the weights provide relevant information about the pertinence of each edge (and their directions, accounted by SHD). This is particularly striking in the non-linear case (Figure 13), where Consensus_05 consistently improves both scores, outperforming the individual approaches. This also demonstrates the flexibility of the consensus since dataset in the non-linear setting are generated through different functional form of the causal relationship (i.e. the function f). Concerning the density of the learned graphs (given by the dots' sizes in both figures), the thresholded consensus graph shows a density consistent with that of the ground truth DAG (the empty dot).

Additional extensive controlled simulation experiments were performed (results not shown) by varying several key parameters: number of nodes, sample size, noise coefficient, noise distribution, functional form of the causal relationship, data discretization. In all settings, we were able to reach similar conclusions as the ones shown here. Overall, these results support the relevance of our proposed consensus approach and provide motivation for ensemble causal discovery.

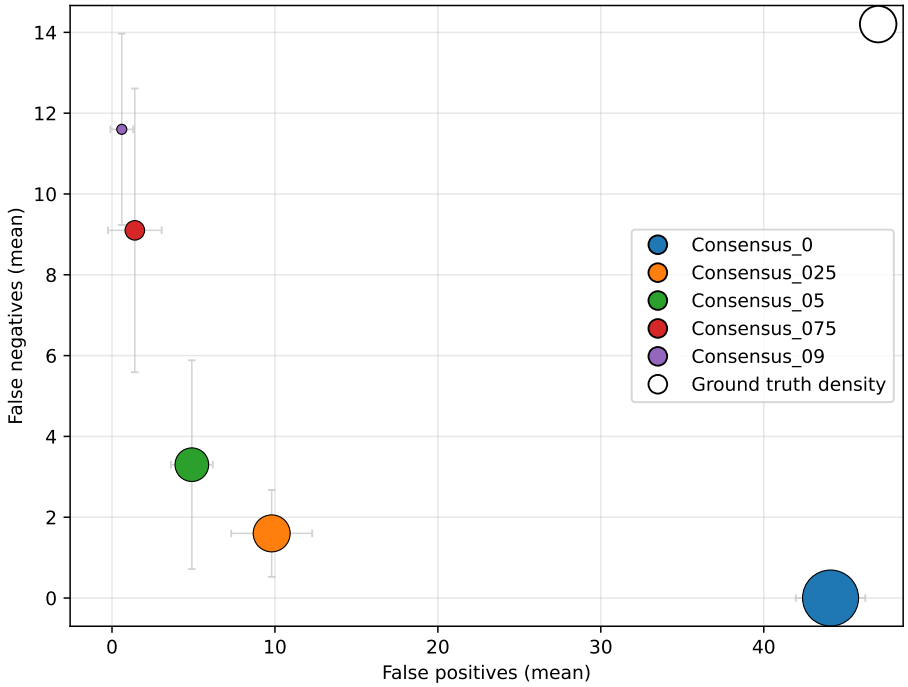


Fig. 14. False positives and False negatives for different threshold consensus. 10 independent datasets of 1000 samples each were randomly generated from 8-node causal graphs, using linear causal relationship ($f(x) = x$), and independent centered Gaussian noise with standard deviation $\sigma = 0.1$. Filled dots represent the mean score over the 10 runs for each method, and error bars represent the corresponding standard deviations. The size of each filled dot is proportional to the average density of the consensus. The empty dot in the top-right corner provides a measure of the average density of the ground truth, as a reference.