Découverte causale sur des jeux de données classiques et temporels. Application à des modèles biologiques

Franck Simon

Thèse de Doctorat de l'université de Sorbonne Université

Ecole doctorale n°130 : EDITE de Paris
Ecole Doctorale Informatique, Télécommunications et Electronique
Spécialité : Sciences et technologies de l'information et de la communication
Thématique : Sciences de l'information et sciences du vivant

Découverte causale sur des jeux de données classiques et temporels Application à des modèles biologiques

Présentée par **Franck SIMON**
Pour obtenir le grade de Docteur de l'université de Sorbonne Université

Réalisée à l'**Institut Curie, UMR 168,**
Laboratoire Physico-Chimie Curie
Sous la direction du **Dr Hervé ISAMBERT**

Soutenue publiquement le **1er décembre 2023**
devant le jury composé de :

M. Habib MEHREZ	Sorbonne Université
M. Eric GAUSSIER	Université Grenoble Alpes
M. Benno SCHWIKOWSKI	Institut Pasteur
M. Charles ASSAAD	EasyVista
Mme Marie Elise MERCUI	Sorbonne Université
Mme Christine MANTECON	Sorbonne Université
M. Hervé ISAMBERT	Institut Curie

Sommaire

Remerciements

En préalable à ce rapport, je tiens à remercier les organismes et les personnes qui m'ont donné l'opportunité de travailler pendant ces 3 années au sein de l'Institut Curie.

Si en premier lieu, il y a bien sûr le Dr Hervé Isambert, professeur et chef d'équipe qui m'a accueilli dans son service, je tiens également à remercier mon responsable administratif au sein de l'Institut Curie, M. Fabrice Demarthon ainsi que mes collègues :

- M. Vincent Cabeli
- M. Honghao Li
- M. Marcel Ribeiro-Dantas
- Melle Louise Dupuis
- Melle Liza Hettal
- Melle Orianne Debeaupuis
- Melle Tiziana Tocci
- M. Nikita Lagrange
- M. Aymeric Monchi
- M. Louis Montagne
- M. Gael Simon

pour leurs conseils et leur aide tout au long de mon parcours.

Outre les membres de l'équipe, je souhaite également remercier l'ensemble des personnes avec lesquelles nous avons mis en place des collaborations et travaillé sur des projets transverses :

- Mme Maria Colomba Comes
- Mme Arianna Mencattini
- Mme Maria Carla Parrini
- M. Eugenio Martinelli
- Mme Hélène Salmon
- M. Abbas Abdenour
- Mme Hélène Moreau
- Melle Zoé Fusilier
- Mme Christel Goudot

Des remerciements spéciaux pour les personnes qui m'ont donné la chance de m'orienter vers la recherche scientifique :

- M. Martin Weigt
- Mme Alessandra Carbone
- M. Denis Thieffry

Enfin, des remerciements sincères aux personnes qui m'ont accompagnées dans cette démarche de doctorat, aux membres de mon jury qui ont pris de le temps d'apprécier mon manuscrit et m'ont donné l'opportunité de soutenir ce mémoire de thèse :

- M. Habib Mehrez
- M. Eric Gaussier
- M. Benno Schwikowski
- M. Pierre-Henri Wuillemin
- M. Charles Assaad
- Mme Marie-Elise Mercui
- Mme Christine Mantecon

Résumé / Abstract

Français

Titre

Découverte causale sur des jeux de données classiques et temporels
Application à des modèles biologiques

Résumé

Cette thèse porte sur le domaine de la découverte causale, c'est-à-dire la construction de graphes causaux à partir de données observées, et en particulier, la découverte causale temporelle et la reconstruction de larges réseaux de régulation de gènes. Après un bref historique, ce mémoire introduit les principaux concepts, hypothèses et théorèmes aux fondements des graphes causaux ainsi que les deux grandes approches : à base de scores et à base de contraintes. La méthode MIIC (Multivariate Information-based Inductive Causation), développée au sein de notre laboratoire est ensuite décrite avec ses dernières améliorations : Interpretable MIIC. Les problématiques et solutions mises en œuvre pour construire une version temporelle (tMIIC) sont exposées ainsi que les benchmarks traduisant les avantages de tMIIC sur d'autres méthodes de l'état de l'art. L'application à des séquences d'images prises au microscope d'un environnement tumoral reconstitué sur des micro-puces permet d'illustrer les capacités de tMIIC à retrouver, uniquement à partir des données, des relations connues et nouvelles. Enfin, cette thèse introduit l'utilisation d'un a priori de conséquence pour appliquer la découverte causale à la reconstruction de réseaux de régulation de gènes. En posant l'hypothèse que tous les gènes, hormis les facteurs de transcription, sont des conséquences, il devient possible de reconstruire des graphes avec des milliers de gènes. La capacité à identifier des facteurs de transcription clés de novo est illustrée par une application à des données de séquençage d'ARN en cellules uniques avec identification de deux facteurs de transcription susceptibles d'être impliqués dans le processus biologique d'intérêt.

Mots-clés

- Causalité
- Découverte causale
- Apprentissage automatique
- Séries temporelles
- Séries chronologiques
- Réseau de régulation génique
- Réseau de régulation de gènes

English

Title

Causal discovery on classical and temporal datasets
Application to biological models

Summary

This thesis focuses on the field of causal discovery : the construction of causal graphs from observational data, and in particular, temporal causal discovery and the reconstruction of large gene regulatory networks. After a brief history, this thesis introduces the main concepts, hypotheses and theorems underlying causal graphs as well as the two main approaches : score-based and constraint-based methods. The MIIC (Multivariate Information-based Inductive Causation) method, developed in our laboratory, is then described with its latest improvements : Interpretable MIIC. The issues and solutions implemented to construct a temporal version (tMIIC) are presented as well as benchmarks reflecting the advantages of tMIIC compared to other state-of-the-art methods. The application to sequences of images taken with a microscope of a tumor environment reconstituted on microchips illustrates the capabilities of tMIIC to recover, solely from data, known and new relationships. Finally, this thesis introduces the use of a consequence a priori to apply causal discovery to the reconstruction of gene regulatory networks. By assuming that all genes, except transcription factors, are only consequence genes, it becomes possible to reconstruct graphs with thousands of genes. The ability to identify key transcription factors de novo is illustrated by an application to single cell RNA sequencing data with the discovery of two transcription factors likely to be involved in the biological process of interest.

Keywords

- Causality
- Causal discovery
- Machine learning
- Time series
- Gene regulatory network
- GRN

Introduction

Nous sommes tous à présent conscients que la corrélation n'implique pas la causalité, mais alors comment trouver les relations causales ? Si c'est une bonne chose que cette distinction soit intégrée dans la culture scientifique moderne, la plupart des études dont l'objectif est de découvrir des mécanismes de causalité se basent encore sur la corrélation comme principal outil pour y parvenir.

La nouvelle science de la causalité essaye de nous réconcilier avec cet objectif, en définissant formellement comment représenter les relations causales, comment les mesurer, et surtout, donner les conditions nécessaires pour les découvrir. La première question sur la façon de représenter les relations causales a probablement trouvé sa meilleure réponse dans la théorie des diagrammes de causalité principalement développée par Judea Pearl [1-3]. Un diagramme causal est un réseau bayésien : un graphe orienté acyclique qui encode les indépendances conditionnelles entre variables aléatoires représentées par les nœuds, avec la dimension causale transcrite par la direction des arêtes. A partir de ces graphiques, on peut déduire des réponses à des questions fondamentalement causales comme "quel est l'effet de ce traitement sur cette population ?", ou encore "et si cette population avait reçu ce traitement ?".

L'équipe du Dr Isambert, au sein du laboratoire qui m'emploie, travaille principalement dans le domaine de la découverte de graphes causaux, qui vise à reconstruire ces modèles graphiques uniquement à partir de données d'observation. Le défi de la découverte causale réside dans la conservation des liens directs qui reflètent une certaine compréhension de la nature, du processus de génération des données, tout en rejetant les fausses interactions qui sont des conséquences indirectes des relations réelles. Dans les bonnes conditions, on sait que l'on peut apprendre le graphe causal jusqu'à un graphique d'équivalence uniquement à partir du modèle de dépendances et d'indépendances trouvé dans les données, sans aucune intervention.

L'équipe se concentre sur les algorithmes basés sur des contraintes en général et MIIC en particulier, une approche de la théorie de l'information combinant des éléments provenant à la fois des méthodes basées sur les contraintes et de celles basées sur les scores. Là où les méthodes classiques reposent sur des tests fréquentistes d'indépendance et un paramètre α pour le seuil de valeur-p, MIIC estime l'indépendance par rapport aux données avec le principe du minimum de longueur de description et la distribution du maximum de vraisemblance normalisée.

Au sein de l'équipe, ma contribution principale a été de construire tMIIC, la version temporelle de la méthode MIIC. En effet, si MIIC était apte à traiter des jeux de données avec un ensemble d'échantillons, l'information de temps, même si elle était disponible, ne pouvait être prise en compte pour la reconstruction du graphe causal.

Au delà de tMIIC, les connaissances que j'ai acquises dans le domaine de la causalité m'ont permis de proposer l'intégration d'un a priori de conséquence que nous avons appliqué à la reconstruction de graphes de régulation de gènes et ce projet, même s'il devra faire l'objet de travaux futurs pour être pleinement finalisé, permet d'ores et déjà des découvertes importantes pour identifier des gènes d'intérêt.

1 La causalité

1.1 Définition usuelle

Le concept de causalité, que l'on appelle communément relation de cause à effet, peut être décrit comme l'influence par laquelle un événement, un processus, un état ou un objet (une cause) contribue à la production d'un autre événement, processus, état ou objet (un effet) où la cause est en partie responsable de l'effet, et l'effet dépend en partie de la cause. En général, un processus a de nombreuses causes, qui sont également considérées comme des facteurs causaux, et toutes se situent dans son passé. Un effet peut à son tour être la cause ou le facteur causal de nombreux autres effets, qui se situent tous dans son avenir.

1.2 Bref historique

1.2.1 Dans l'antiquité

Cependant, cette notion de causalité à fortement évolué au travers des âges, tant du point de vue philosophique que scientifique. L'humanité a conscience de la causalité depuis des temps très reculés puisque l'on peut la retrouver dans les anciennes écritures hindoues : « La cause est l'effet caché, l'effet est la cause révélée » [4] ou dans des textes de la Grèce ancienne.

C'est probablement Platon qui a le premier énoncé le principe de causalité : "tout ce qui devient ou change doit le faire en raison d'une cause ; car rien ne peut arriver sans cause". Platon a souligné l'importance causale des causes formelles, rien ne peut être s'il n'y a pas un modèle immuable de causes formelles dont le phénomène observable individuel n'est qu'une simple apparence [5].

Aristote a ensuite introduit les principes des quatre causes : matérielle (la matière qui constitue une chose), formelle (l'essence de cette chose), motrice (le principe de changement) et cause finale (l'usage de la chose) [6].

1.2.2 Causalité et régularité

Si de nombreux auteurs ont ensuite fait évoluer la notion de causalité aux cours des siècles, c'est certainement Hume qui a eu la contribution la plus influente au 18ème siècle dans le domaine de recherche de cette thèse, la découverte de relations causales à partir d'un jeu de données. S'opposant au rationalisme, qui suppose une connaissance a priori de la causalité, Hume soutient que la raison pure ne peut à elle seule prouver la réalité de la causalité et au lieu de cela, il a fait appel à la coutume et à l'habitude mentale, observant que toute connaissance humaine découle uniquement de notre expérience sur un grand nombre de cas similaires.

Aujourd'hui, l'idée de Hume selon laquelle la régularité ou la conjonction constante est une condition nécessaire à la causalité est généralement acceptée. Ce point de vue semble s'accorder avec notre bon sens, nous nous attendons à ce que des causes similaires aient des effets similaires, mais Hume soutenait que la régularité est aussi une

condition suffisante de la causalité. Ce point de vue a été facilement démontré comme faux car il existe de nombreux exemples de conjonctions constantes, comme le jour suivant la nuit, qui ne sont pas des relations causales [5].

1.2.3 Diagrammes de chemins

L'analyse de trajectoires a été développée vers 1918 par le généticien Sewall Wright [7] et constitue l'ancêtre des graphes causaux actuels [1]. Cette méthode tente de mesurer l'influence directe de chaque corrélation et de déterminer dans quelle mesure la variation d'un effet donné est déterminée par chaque cause particulière. Dans ce but, des diagrammes de variables reliées par des flèches sont construits, montrant les différentes corrélations au sein du système. Sur la base de ces diagrammes et des corrélations observées entre les variables, des équations sont construites et alors résolues.

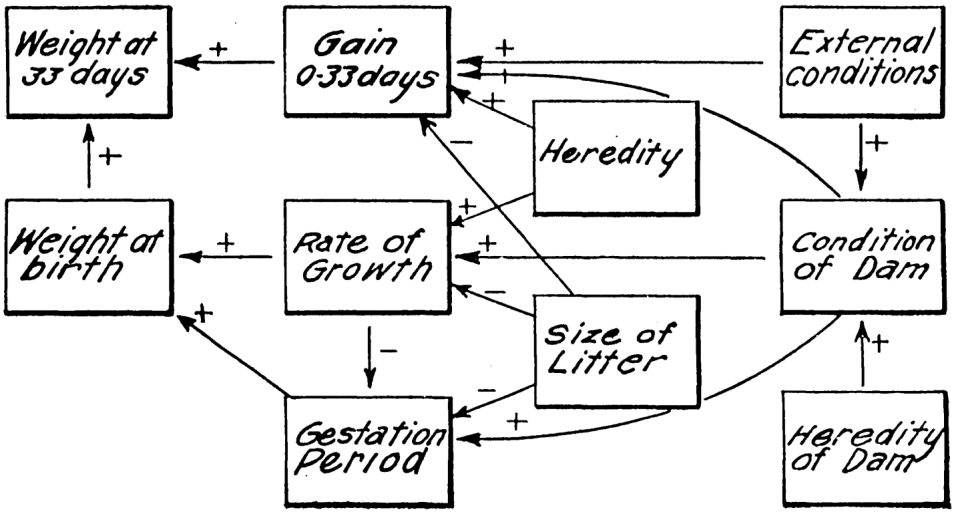


FIGURE 1.1 – **Diagramme de chemin** illustrant les interrelations entre les facteurs qui déterminent le poids des cochons d'Inde à la naissance et au sevrage (33 jours).

Si ces modèles étaient initialement confinés à des équations linéaires à paramètres fixes, ils ont servi de base aux développements modernes qui ont étendu les modèles graphiques à l'analyse non paramétrique et ont ainsi atteint une généralité et une flexibilité qui ont transformé l'analyse causale en informatique, en épidémiologie et en sciences sociales.

1.2.4 Essais contrôlés randomisés

Dans l'étude de cause à effet d'un traitement, la mise en œuvre des essais contrôlés randomisés (ECR) permet de fournir des preuves convaincantes que le traitement à l'étude a un effet sur la santé. Le premier essai contrôlé randomisé publié en médecine est apparu dans l'article de 1948 intitulé "Traitement à la streptomycine de la tuberculose pulmonaire", qui décrivait une enquête du Medical Research Council [8, 9]. L'un des auteurs de cet article était Austin Bradford Hill, à qui l'on attribue la conception de l'ECR moderne.

Un ECR en recherche clinique compare généralement un nouveau traitement proposé à un soin standard existant, ceux-ci sont alors appelés respectivement les traitements « expérimental » et « de contrôle ». Lorsqu'aucun traitement de contrôle n'est

disponible, un placebo peut être utilisé dans le groupe témoin afin que les participants ne connaissent pas leurs allocations de traitement. Idéalement, ce principe d'aveuglement est également étendu autant que possible à d'autres parties, y compris les chercheurs, les techniciens, les analystes de données et les évaluateurs. Une mise en aveugle efficace isole expérimentalement les effets physiologiques des traitements de diverses sources psychologiques de biais. Le caractère aléatoire dans l'affectation des participants aux traitements réduit le biais de sélection et le biais d'attribution, en équilibrant les facteurs pronostiques connus et inconnus, dans l'attribution des traitements. L'aveuglement réduit les autres formes de préjugés de l'expérimentateur et du sujet.

Un ECR bien aveuglé est considéré comme l'étalon-or pour les essais cliniques. Les ECR en aveugle sont couramment utilisés pour tester l'efficacité des interventions médicales et peuvent en outre fournir des informations sur les effets indésirables, tels que les réactions aux médicaments.

1.2.5 Causalité probabiliste

Augmentation de probabilité

Plutôt que d'essayer d'analyser la causalité selon les théories de la régularité de la causalité (cf 1.2.2), où les mécanismes causaux sont une succession de schémas invariables, l'idée centrale derrière les théories probabilistes de la causalité est que les causes modifient la probabilité de leurs effets : un effet peut toujours se produire en l'absence d'une cause ou ne pas se produire en sa présence. Ainsi, le tabagisme est une cause de cancer du poumon, non pas parce que tous les fumeurs développent un cancer du poumon, mais parce que les fumeurs sont plus susceptibles de développer un cancer du poumon que les non-fumeurs. Cela est tout à fait cohérent avec le fait qu'il y a des fumeurs qui évitent le cancer du poumon et des non-fumeurs qui y succombent.

Cette idée centrale selon laquelle les causes augmentent la probabilité de leurs effets peut être exprimée formellement en utilisant la probabilité conditionnelle. C augmente la probabilité de E seulement lorsque :

$$P(E|C) > P(E) \tag{1.1}$$

En d'autres termes, la probabilité que E se produise, étant donné que C se produit, est supérieure à la probabilité inconditionnelle que E se produise. Alternativement, nous pourrions dire que C augmente la probabilité de E seulement lorsque :

$$P(E|C) > P(E|\overline{C}) \tag{1.2}$$

la probabilité que E se produise, étant donné que C se produit, est supérieure à la probabilité que E se produise, étant donné que C ne se produit pas. Ces deux formulations s'avèrent équivalentes en ce sens que l'inégalité 1.1 est valide lorsque 1.2 l'est.

Hans Reichenbach fut l'un des premiers, en 1956, à relier causalité avec probabilité [10]. Dans son ouvrage, la direction du temps, Reichenbach présente la première théorie probabiliste entièrement développée de la causalité [11].

Filtrage

Reichenbach a introduit la terminologie de filtrage pour décrire un type particulier de relation probabiliste. Si $P(E|A,C) = P(E|C)$, alors on dit que C sépare A de E. Lorsque $P(A,C) > 0$, cette égalité est équivalente à $P(A,E|C) = P(A|C) \times P(E|C)$, c'est-à-dire que A et E sont probabilistiquement indépendants conditionnellement à C.

Reichenbach a reconnu qu'il y avait deux types de structures causales dans lesquelles C sépare typiquement A de E. La première se produit lorsque A cause C, qui à son tour cause E, et il n'y a pas d'autre voie ou processus par lequel A affecte E et est illustré à la figure 1.2. Dans ce cas, Reichenbach dit que C est causalement entre A et E. On pourrait dire que C est une cause intermédiaire entre A et E, ou que C est une cause proximale de E et A une cause distale de E.

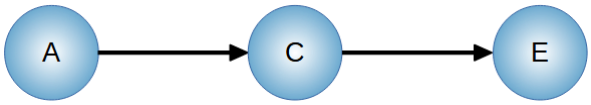


FIGURE 1.2 – **Premier cas de filtrage**, C est causalement entre A et E.

Le deuxième type de cas qui produit un filtrage arrive lorsque C est une cause commune de A et E (Fig. 1.3). Un exemple de la vie réelle peut être celui de la météo : une baisse de la pression atmosphérique (C) entraîne à la fois une baisse du niveau de mercure dans un baromètre (A) et une tempête (E). La pression atmosphérique masquera donc le lien entre la lecture du baromètre et la météo : étant donné que la pression atmosphérique a chuté, la lecture du baromètre ne fait aucune différence pour la probabilité qu'une tempête se produise.

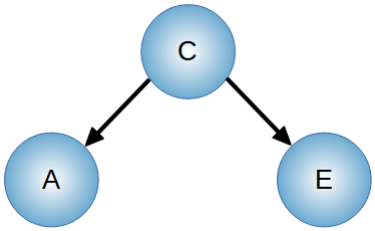


FIGURE 1.3 – **Second cas de filtrage**, C est une cause commune de A et E.

Corrélations fallacieuses

Reichenbach a utilisé la notion de filtrage pour résoudre le problème des fausses corrélations. Dans notre exemple, alors qu'une baisse dans la colonne de mercure (A) augmente globalement la probabilité d'une tempête (E), elle n'augmente pas la probabilité d'une tempête lorsque nous conditionnons sur la pression atmosphérique. Autrement dit, si A et E sont faussement corrélés, alors A sera séparé de E par une cause commune. Plus précisément, supposons que C_t et $E_{t'}$ soient des événements qui se produisent aux instants t et t' respectivement. Alors :

C_t est une cause de $E_{t'}$ si et seulement si :
1) $t < t'$
2) $P(E_{t'}|C_t) > P(E_{t'}|\overline{C_t})$ (1.3)
3) Il n'y a pas d'autre événement $B_{t''}$, se produisant à un instant t'' antérieur ou simultané à t, qui sépare $E_{t'}$ de C_t.

Le principe de cause commune

Reichenbach a également formulé un principe qu'il a surnommé le «principe de cause commune» (PCC). Supposons que les événements A et B soient corrélés positivement, c'est-à-dire que :
1) $P(A,B) > P(A) \times P(B)$
mais sans que ni A, ni B ne soient cause l'un de l'autre. Reichenbach a soutenu qu'il y aura une cause commune, C, de A et B, satisfaisant les conditions suivantes :
2) $0 < P(C) < 1$
3) $P(A,B|C) = P(A|C) \times P(B|C)$ (1.4)
4) $P(A,B|\overline{C}) = P(A|\overline{C}) \times P(B|\overline{C})$
5) $P(A|C) > P(A|\overline{C})$
6) $P(B|C) > P(B|\overline{C})$

Lorsque les événements A, B et C satisfont ces conditions, on dit qu'ils forment une fourche conjonctive. 5 et 6 découlent du fait que C est une cause de A et une cause de B. Les conditions 2 et 3 stipulent que C et $\overline{C}$ filtrent A de B.

Les conditions 2 à 6 impliquent mathématiquement l'inéquation 1. Reichenbach dit que la cause commune explique la corrélation entre A et B. L'idée est que les corrélations probabilistes qui ne sont pas le résultat d'un événement en provoquant un autre sont finalement dérivées de corrélations probabilistes qui résultent d'une cause commune.

Dans la définition de Reichenbach de la causalité, l'inégalité $P(E_{t'}|C_t) > P(E_{t'}|\overline{C_t})$ est nécessaire, mais pas suffisante, pour la causalité. Malheureusement, les causes communes peuvent également donner lieu à des cas où cette inégalité n'est pas non plus nécessaire à la causalité, comme dans le cas du paradoxe de Simpson, qui peut inverser les inégalités probabilistes.

1.2.6 Le modèle contrefactuel

La causalité, définie en termes contrefactuels, revient à dire que, pour prouver qu'un évènement en cause un autre, il faut prouver que, si le premier évènement ne s'était pas produit, le second n'existerait pas [12].

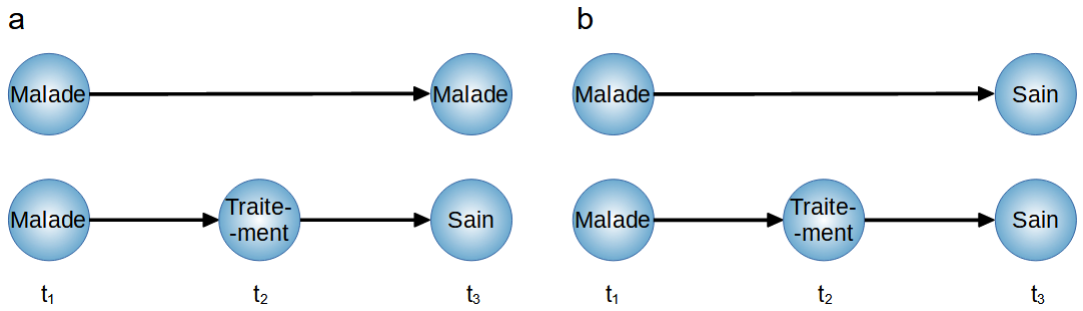


FIGURE 1.4 – **a** Sans traitement, la maladie n'est pas guérie, ce qui indique un effet causal du traitement. **b** Le rétablissement se produit avec ou sans traitement, le traitement n'est pas la cause du rétablissement.

Cependant, la dépendance contrefactuelle seule ne pourrait pas former une règle de causalité, parce qu'elle est suffisante mais non nécessaire pour la causalité. Deux événements peuvent être liés causalement sans que l'un dépende contrefactuellement de l'autre, par exemple lorsqu'il existe des alternatives qui peuvent entraîner l'effet en l'absence de la cause.

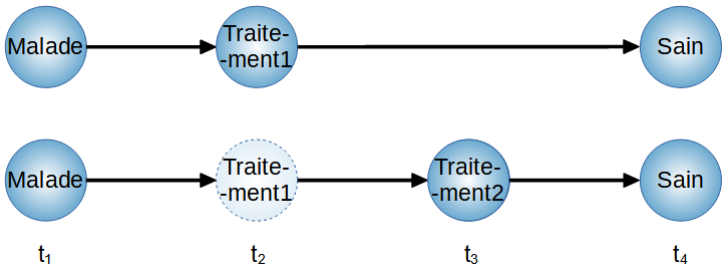


FIGURE 1.5 – **Exemple de non nécessité** Si le délai de prescription du traitement 1 est dépassé, le traitement 2 est prescrit et permet le rétablissement sans que le traitement 1 n'ait été nécessaire.

Bien que David Lewis ait amendé sa théorie contrefactuelle en introduisant les notions d'altération (un événement identique à un autre sauf qu'il se produit à un moment, à un endroit, à un lieu légèrement différent ou d'une manière légèrement différente) et d'influence contrefactuelle (un événement en provoque un autre si et seulement s'il existe une chaîne d'influences progressives (altérations) ramenant du dernier au premier), la théorie contrefactuelle est cependant restée la cible de critiques [10]. La notion de contrefactualité sera cependant reprise dans les modèles causaux structurels combinant les caractéristiques des modèles d'équations structurelles, les modèles graphiques et l'explication contrefactuelle de la causalité.

1.2.7 La théorie interventionnelle

La théorie interventionnelle de la causalité a été introduite par Woodward [13] alors que Pearl [1] et Spirtes [14] se concentrent sur les contraintes formelles qu'une intervention impose à un système de variables.

Pour illustrer le principe d'une intervention, nous allons utiliser l'exemple de la figure 1.6 où nous voulons évaluer si la lecture du niveau de mercure d'un baromètre peut être la cause des tempêtes. Pour cela, nous réalisons donc une intervention sur

Baromètre en introduisant une variable *Intervention* qui nous permet de fixer la valeur de *Baromètre* quelle que soit la pression atmosphérique : nous supprimons toute influence possible sur la variable *Baromètre* de sorte que sa valeur soit entièrement fixée par l'intervention.

Pour que l'intervention soit valide, il faut également que :

- l'intervention ne cause pas *Tempête* sans passer par *Baromètre* ;
- l'intervention ne doit pas être causée par toute cause affectant *Tempête* via une route qui ne passe pas par *Baromètre* ;
- l'intervention doit laisser inchangées les valeurs prises par toute cause de *Tempête* exceptées celles sur le chemin direct de *Baromètre* à *Tempête*.

Si ces exigences sont respectées, on pourra conclure que *Baromètre* cause *Tempête* lorsque la modification par intervention de *Baromètre* modifie de manière régulière la valeur de *Tempête*.

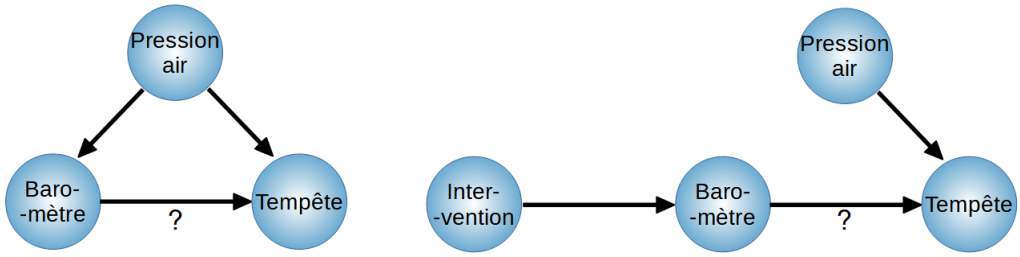


FIGURE 1.6 – **Exemple d'intervention.** Pour évaluer l'arête *Baromètre* → *Tempête*, nous réalisons une intervention sur *Baromètre*.

Une telle intervention est notée $p(Y|do(X = x))$ et est la base du do-calculus. Remarquons que $p(Y|X)$ et $p(Y|do(X = x))$ ne sont pas identiques : le premier est observationnel tandis que le second est interventionnel. Sur notre exemple, soit X la lecture sur le baromètre et Y l'occurrence d'une tempête, si nous les observons ensemble chaque jour et notons leurs valeurs, $p(Y|X)$ montrerait une forte relation entre les deux : si le baromètre mesure une basse pression atmosphérique, le temps est souvent mauvais, tandis que s'il mesure une haute pression, le temps est meilleur. La distribution $p(Y|do(X = x))$ ne sera cependant pas réellement dépendante de la valeur X : on ne peut pas changer le temps en réglant la lecture sur un baromètre. Même si $p(Y|X)$ peut être utilisé pour prédire la valeur d'une variable en mesurant l'autre, il n'informe pas sur la relation fonctionnelle entre les deux. Pour cela, nous devons monter d'un échelon sur l'échelle de la causalité, en utilisant le do-calculus.

1.2.8 L'échelle de causalité

Pour synthétiser cette partie sur l'historique et quelques grandes étapes dans le domaine de la causalité, il est possible de reprendre l'illustration de la figure 1.7 représentant les 3 niveaux de l'investigation causale.

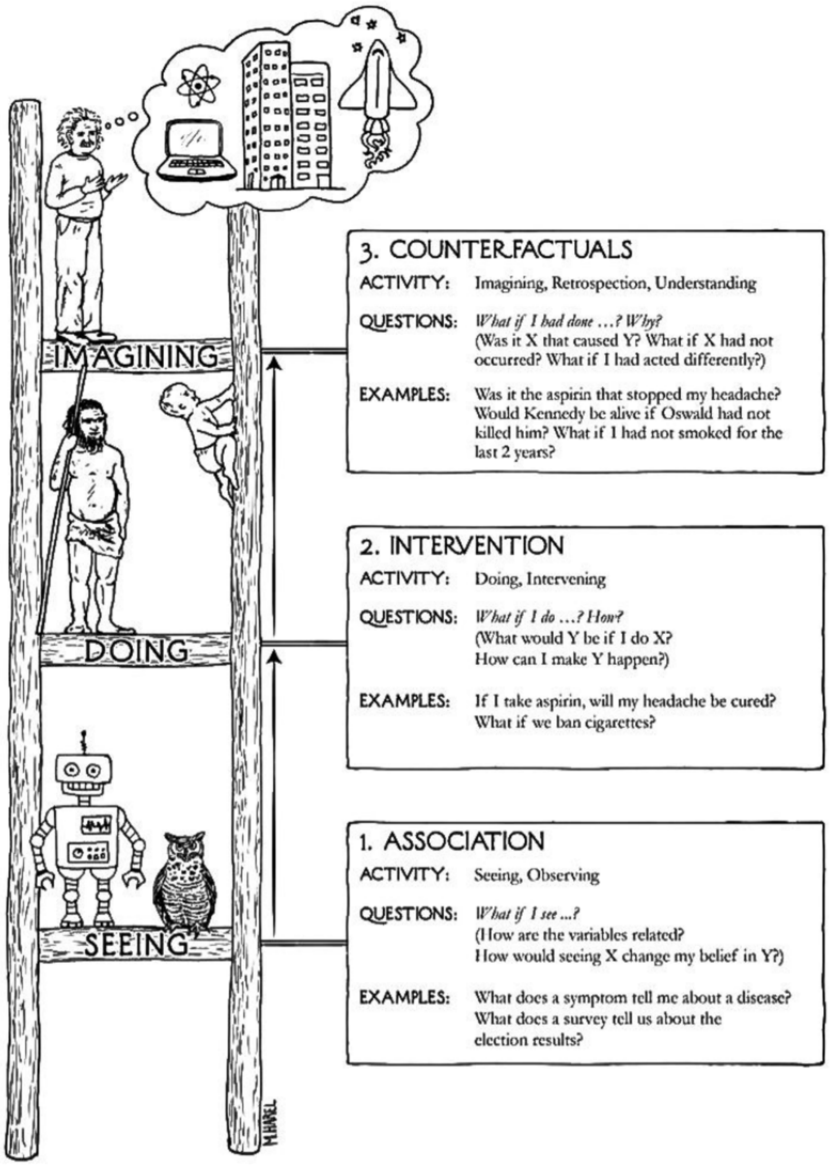


FIGURE 1.7 – **L'échelle de causalité.** (extrait de "The book of why" de Pearl and Mackenzie, 2018)

De nombreux animaux et la plupart des algorithmes d'apprentissage automatique actuels sont au premier échelon de l'échelle. Ils apprennent par simple observation, et travaillent donc par association. Un exemple de question d'association est : si je vois A se produire, est-ce que cela augmente la chance de voir B arriver ? Il n'y a pas d'action ou d'intervention ici. Vous disposez de quelques observations et vous vous demandez ce que vous pouvez en déduire.

Au deuxième échelon, nous trouvons les premiers humains et quelques autres animaux qui ont réussi à développer des outils et un raisonnement stratégique. On peut se demander si je fais des exercices tôt le matin, est-ce que cela me fera me sentir mieux tout au long de ma journée de travail ? Il y a une intervention ici, il y a de l'action. La personne peut essayer ceci et apprendre de l'expérience. Les essais contrôlés randomisés sont à cet échelon. Un détail important à garder à l'esprit est que nous parlons ici

de l'avenir et d'une réalité qui pourrait devenir réelle.

Ce détail est l'une des principales différences entre les échelons 2 et 3. Bien qu'à l'échelon 1 on ne puisse que voir, et qu'à l'échelon 2 on puisse également faire, intervenir, à l'échelon 3, on peut seulement imaginer : la chose que nous imaginons ne se réalisera jamais.

Pour mieux comprendre cela, imaginons une expérience qui tente d'identifier si le traitement A a un effet causal sur une certaine mesure de la récupération par rapport au traitement habituel (appelons-le traitement B). Les individus seront assignés au hasard au groupe 1 (traitement A) ou au groupe 2 (traitement B). À la fin, nous pouvons calculer l'effet moyen du traitement et, dans ce cas hypothétique, nous avons constaté qu'il y a un effet causal du traitement A dans la guérison des patients : ils récupèrent plus rapidement, par rapport au traitement B. Comment serait-il possible d'aborder cette situation d'une perspective à l'échelon 3 ? On pourrait demander : je vois que le patient numéro 1523 a reçu le traitement B et a mis X jours à se rétablir. Quelle serait la différence si ce même patient avait reçu le traitement A ? Il n'y a pas d'expérience possible qui nous permettrait de tester cela de manière empirique. Le patient 1523 a reçu le traitement B et le seul moyen de changer serait de voyager dans le temps, ce qui est impossible avec la technologie actuelle comme nous le savons. L'examen des questions de niveau 3 est beaucoup plus compliqué que celui de questions de niveau 2 et il est donc très important d'exposer que ces requêtes sont différentes. Plus on monte dans l'échelle de causalité, plus il faut d'hypothèses pour étudier l'événement.

1.3 Les graphes causaux

Dans cette section va être exposée l'utilisation de graphes pour représenter la causalité, avec tout d'abord, un exemple intuitif d'un diagramme de causalité pour ensuite montrer que de tels graphes peuvent décrire naturellement des manières communes d'inférer la causalité, que ce soit par des expériences ou à partir de données d'observation. Enfin, nous donnons les notations formelles et les définitions du cadre des graphes causaux.

1.3.1 Intuition des graphes causaux

Considérons une situation familière dans laquelle notre intuition peut être représentée par une relation causale (Fig. 1.8). Supposons qu'il y ait deux causes qui pourraient être à l'origine de la *Panne* de la voiture, *Batterie* (batterie déchargée) et *Huile* (niveau d'huile moteur bas) et que nous essayons de faire un diagnostic avant d'intervenir. Dans le schéma illustrant leurs relations visuellement, les deux causes considérées, *Batterie* et *Huile* sont représentées en tant que parents du nœud *Panne*. Comme il n'y a aucune raison de penser que les deux causes considérées soient liées, nous le représentons en ne dessinant pas d'arête entre elles. Nous incluons un quatrième nœud *Phares* qui correspond à un autre constat : les phares ne s'allument pas. Nous savons que les phares ne dépendent pas du niveau d'huile mais ils ont besoin de batterie pour fonctionner, et le nœud *Phares* est donc lié uniquement au nœud *Batterie*. Nous savons aussi qu'habituellement, si les phares ne s'allument pas, la voiture ne démarrera probablement pas non plus. Nous représentons cette association par un lien pointillé, reflétant une corrélation, mais pas une relation causale : l'interaction indirecte n'existe

qu'à cause de l'ancêtre commun *Batterie* mais ne nous informe pas d'une relation fonctionnelle (elle ne serait donc pas incluse dans le diagramme causal). Cela peut nous aider à deviner l'origine de la panne (*Batterie* ou *Huile*), cependant, la réparation des phares n'aidera pas à démarrer la voiture.

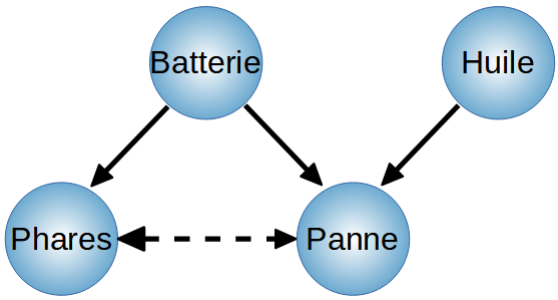


FIGURE 1.8 – **Exemple de graphe causal.** Graphe de diagnostic d'une voiture en panne.

Dans cet exemple de la figure 1.8, nous dessinons le graphe à partir de connaissances pré-existantes, mais que peut-on faire lorsque ces connaissances ne sont pas disponibles ? C'est le domaine de l'inférence causale, qui vise à découvrir les effets de causalité soit par l'expérimentation ou l'observation passive du système.

1.3.2 Approches d'inférence causale utilisant les graphes causaux

L'étalon-or de l'inférence causale est l'essai contrôlé randomisé, où une population homogène se voit attribuée au hasard soit un traitement, soit un placebo. Définissons Y le résultat de l'essai pour chaque patient, qui peut être positif ou négatif, nous voulons savoir dans quelle mesure le résultat dépend du traitement, noté X, par opposition à d'autres facteurs externes qui sont tous regroupés dans le nœud Z. Formellement, on peut répondre à cette question en comparant $p(Y|do(X = traitement))$ et $p(Y|do(X = placebo))$. Le graphe causal d'un véritable essai randomisé est illustré à la figure 1.9.

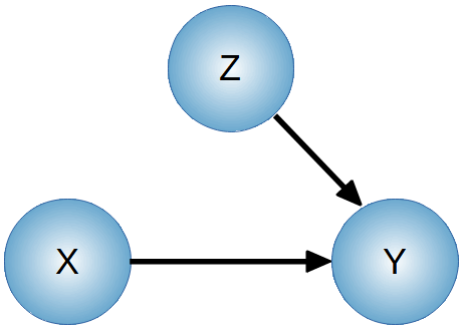


FIGURE 1.9 – **Essai contrôlé randomisé** où X est le traitement, Y le résultat et Z représente les facteurs externes.

Nous pouvons voir sur le diagramme causal que l'effet causal de X sur Y est direct, il n'est pas affecté par le reste du graphique. Dans ce contexte, l'attribution aléatoire de X est une sorte d'intervention, et $p(Y|do(X))$ peut être directement observé à partir des données comme $p(Y|X)$. Ce type d'expérience est généralement fiable à condition que le traitement soit assigné vraiment au hasard, mais il comporte cependant des

inconvénients. Tout d'abord, il doit être effectué pour chaque X pour lequel nous voulons savoir l'effet, et il peut être trop long ou trop difficile de recruter suffisamment de participants. Ensuite, il peut être contraire à l'éthique lorsque nous soupçonnons que l'interaction est nuisible, pensez par exemple à forcer des sujets d'essai à être exposés à des agents cancérigènes. Enfin, il peut être tout simplement impossible d'intervenir sur la cause potentielle, par exemple, nous ne pouvons pas randomiser la constitution génétique des patients pour étudier la prévalence de certaines maladies.

Pour ces cas, nous pouvons toujours effectuer une inférence causale en observant simplement la cause potentielle, le résultat et tous les facteurs de confusion qui affectent à la fois la cause et l'effet (Fig. 1.10).

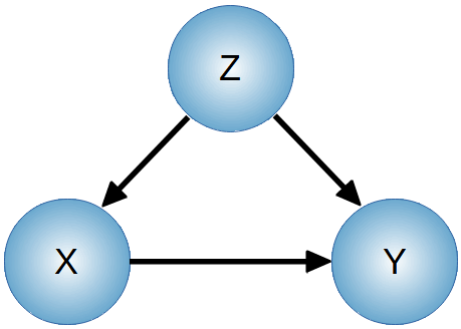


FIGURE 1.10 – **Essai contrôlé randomisé** où X est le traitement, Y le résultat et Z représente les facteurs externes.

Ce graphique résume le principe de plusieurs approches, même celles qui n'ont pas adopté le langage des graphes causaux et du do-calculus. Les procédures d'appariement, par exemple, font exactement les mêmes hypothèses pour estimer l'effet de X sur Y en prenant en compte l'effet de Z. Leur objectif est de réduire le biais d'affectation pour le "traitement" X et imiter un essai contrôlé randomisé en créant des échantillons appariés sur Z, essentiellement en supprimant l'arête $Z \to X$ [15, 16]. La figure 1.10 est également le cadre typique où l'on peut modéliser simplement Y à partir de X tout en ajustant sur Z. C'est l'approche adoptée par des études associatives à l'échelle du génome, qui tentent de mesurer l'effet de milliers de gènes X sur l'apparition d'une maladie Y en ajustant certaines composantes principales Z pour modéliser les différences d'ascendance entre cas et témoins [17].

Toutes ces méthodes sont théoriquement valables mais sont souvent critiquées pour leur prédisposition à donner des résultats biaisés. Selon Pearl, ces défauts proviennent principalement du fait que Z doit contenir autant de covariables que possible, pour ajuster avec toute l'information disponible [1, p. 350, 351]. En ignorant les "conditions d'ignorabilité forte" pour une variable à inclure dans Z, nous finirons inévitablement par inclure des variables qui ne sont pas des parents mais des enfants de X et Y : $X \to Z \leftarrow Y$, violant les hypothèses et le graphe de la Fig 1.10. Toujours selon Pearl, ce genre d'erreur est beaucoup moins susceptible de se produire lors de l'utilisation du cadre des graphes causaux, car les praticiens sont obligés de modéliser au préalable les interactions, en pensant aux relations causales entre le traitement, le résultat et les covariables.

Comme dernier exemple, nous allons regarder le cas où nous ne pouvons toujours pas intervenir sur la cause potentielle X, et nous savons qu'elle est affectée par des

facteurs de confusion communs de Y, mais nous ne pouvons pas les mesurer, ni les ajuster. Ceci a été notoirement la défense d'éminents statisticiens employés par les compagnies de tabac au moment des premiers rapports établissant un lien entre les cigarettes et le cancer du poumon. L'association ne pouvait être niée, mais ils prétendaient qu'elle pouvait s'expliquer par une cause commune cachée, un facteur génétique par exemple, qui aurait été la cause pour une certaine population, à la fois de vouloir fumer et de développer plus de cancers que la population générale. En 1964, la technologie de séquençage n'était pas encore disponible, et comme les essais contrôlés randomisés étaient hors de question, cet argument était difficile à réfuter et aurait retardé la législation anti-tabac [1, p. 83]. Dans ce cadre, nous pouvons toujours mesurer l'effet de X sur Y si nous mesurons une autre variable "instrumentale" I dont on sait qu'elle a un effet sur X et qu'elle est indépendante des facteurs de confusion latents L (Fig. 1.11).

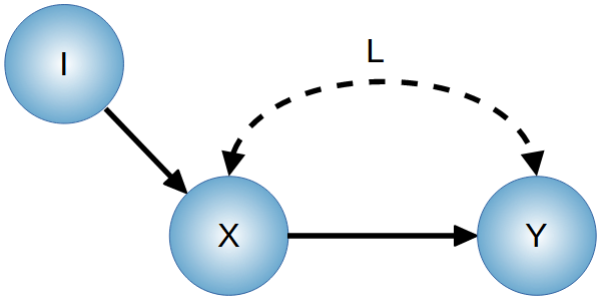


FIGURE 1.11 – **Variable instrumentale** : lorsque les variables de confusion ne sont pas observées, l'effet de X sur Y peut être estimé à partir d'une variable instrumentale I.

En effet, il est possible de voir sur le graphique que toute association mesurée entre I et Y passe nécessairement par X, prouvant l'existence de l'arête dirigée $X \rightarrow Y$. Dans l'exemple pour voir si les cigarettes causent le cancer du poumon, on peut prendre pour I le prix des paquets de cigarettes. Intuitivement, si I est corrélé au nombre de cancers du poumon Y, il est possible d'en déduire l'existence d'un lien causal entre la consommation de tabac et le cancer [18]. Comme pour les variables d'ajustement de l'exemple précédent, modéliser les interactions dans un graphe causal est un moyen de s'assurer que I correspond à une variable instrumentale.

Cette brève discussion met en évidence la remarquable adaptabilité des graphes causaux. Ces modèles intuitifs sont capables de résumer la plupart, sinon toutes les approches qui visent à déduire la causalité. Une fois le graphe causal connu, il est alors possible de dériver des quantités formelles d'effets causaux ainsi que d'appliquer la théorie contre-factuelle en utilisant le do-calculus.

1.3.3 Notations et définitions

Le cadre des graphes causaux traite la causalité comme une propriété statistique, il utilise des éléments de la théorie des graphes et des probabilités. Dans cette section, nous allons passer en revue les notions de base sous-tendant les approches d'inférence causale [14, 19].

En premier lieu, les principales notations qui seront utilisées dans ce document :

- Soit $\mathcal{D}$ un jeu de données composé de V variables $X_1, ..., X_v$. Pour les cas par paires et conditionnels, nous pouvons utiliser X, Y et Z comme variables de $\mathcal{D}$ à la place.
- Chaque variable a une distribution $p(X_i)$, et la distribution conjointe de $\mathcal{D}$ est $p(V)$.
- On note deux variables indépendantes par $X \perp\!\!\!\perp Y$, et conditionnellement indépendantes par rapport à Z par $X \perp\!\!\!\perp Y \mid Z$.
- $\mathcal{D}$ est représenté par un graphe $\mathcal{G}$ où chaque variable est un nœud. Nous notons la vrai graphe causal $\mathcal{G}_c$ et le graphe inféré depuis les données $\mathcal{G}_{Inf}$.
- Si les variables X et Y de $\mathcal{D}$ sont adjacentes dans $\mathcal{G}$, l'arête entre les deux est soit non orientée $X - Y$, orientée $X \rightarrow Y$, $X \leftarrow Y$ ou bi-directionnelle $X \leftrightarrow Y$.
- Les variables qui ont une arête pointant vers X_i sont ses parents et sont notés Pa_i. Les variables vers lesquelles X_i pointe sont ses enfants, notés Ch_i.
- Le squelette de $\mathcal{G}$ est le graphe avec les mêmes adjacences et sans arête orientée.
- Une structure en V est un sous-graphe de trois nœuds où $X \rightarrow Z \leftarrow Y$ avec $X \not\sim Y$.
- Le graphe complet sur les variables V est le squelette où tous les X_i, X_j sont adjacents.

L'objectif de la découverte causale étant de construire un graphe causal uniquement à partir des données d'observation, les relations entre distribution de probabilité et représentation graphique sont au cœur de cette construction. Pour cela, nous avons besoin d'un ensemble d'hypothèses :

Définition 1.1. Suffisance causale : pour chaque paire de variables qui ont leurs valeurs observées dans un ensemble de données, toutes leurs causes communes ont également des observations dans l'ensemble de données. Dit de manière différente, la suffisance revient à exclure la présence de variables latentes. Toutefois, certaines méthodes de découverte causale (par ex FCI, MIIC) n'imposent pas cette hypothèse et le graphe inféré peut contenir des arêtes bi-directionnelles représentant des associations engendrées par des causes communes non observées, c.a.d. des variables latentes.

Définition 1.2. La *d-separation* est le moyen de couper le flux de causalité d'un nœud à un autre :
Un chemin entre deux nœuds X et Y d'un DAG (Directed Acyclic Graph) $\mathcal{G}$ est d-séparé par un ensemble de nœuds Z (qui peut être vide) si et seulement si :

- ce chemin entre X et Y contient une chaîne $i \rightarrow ... \rightarrow m \rightarrow ... \rightarrow j$ ou $i \leftarrow ... \leftarrow ...m \rightarrow ... \rightarrow j$ avec $m \in Z$, ou
- ce chemin entre X et Y contient un collisionneur $i \rightarrow ... \rightarrow m \leftarrow ... \leftarrow j$ tel que $m \notin Z$ et aucun descendant de m ne se trouve dans Z .

L'ensemble Z d-sépare X et Y si et seulement si Z bloque chaque chemin entre X et Y de cette manière.

La découverte causale consistant à construire un graphe compatible avec les relations de dépendances et d'indépendances (conditionnelles) présentes dans les données, le théorème suivant énonce une condition nécessaire et suffisante pour qu'un DAG et une distribution de probabilité soient compatibles.

Théorème 1.1. *Condition de Markov : $\mathcal{G}$ et $p(V)$ satisfont la condition de Markov si et seulement si, compte-tenu de l'ensemble de tous ses parents, un nœud de $\mathcal{G}$ est probabilistiquement indépendant de tous ses non descendants de $\mathcal{G}$.*

Comme plusieurs DAG peuvent représenter le même ensemble d'indépendances conditionnelles et être compatibles avec la même distribution de probabilité, la possibilité de déduire un graphe causal à partir des seules probabilités est limitée. Deux conditions supplémentaires principales ont ainsi été introduites afin de restreindre les graphes considérés à partir d'une distribution de probabilité donnée. La première est la condition de minimalité, qui exige que le graphique ne contienne pas de dépendances non présentes dans les données observationnelles.

Définition 1.3. Un DAG $\mathcal{G}$ compatible avec une distribution de probabilité $p(V)$ est dit satisfaire la condition de minimalité si $p(V)$ n'est pas compatible avec aucun sous-graphe propre de $\mathcal{G}$.

La condition de minimalité n'est cependant pas suffisante pour restreindre l'ensemble des graphes causaux possibles. Pour cela, nous avons besoin de la notion de fidélité.

Définition 1.4. Fidélité : on dit qu'un graphe $\mathcal{G}$ et une distribution de probabilité compatible $p(V)$ sont fidèles l'un à l'autre si toutes et seules les relations d'indépendance conditionnelle vraies dans $p(V)$ sont entraînées par la condition de Markov appliquée à $\mathcal{G}$.

Il est intéressant de remarquer que la notion de minimalité est plus faible que la fidélité dans le sens où la fidélité et les conditions de Markov impliquent ensemble la minimalité, alors que la minimalité et les conditions de Markov n'impliquent pas toujours la fidélité.

La classe des graphes qui sont observationnellement équivalents est appelée la classe d'équivalence de $\mathcal{G}$ [20] :

Définition 1.5. Une classe d'équivalence de Markov est un ensemble de DAG qui encode le même ensemble d'indépendances conditionnelles [21].

Théorème 1.2. *Deux DAG sont dans la même classe d'équivalence si et seulement s'ils ont les mêmes squelettes et les mêmes ensembles de structures en V.*

Dans le contexte de la découverte causale temporelle, d'autres notions sont également intéressantes à introduire :

Définition 1.6. Priorité temporelle : une relation causale entre deux variables est dite satisfaire à la priorité temporelle si elle est orientée de telle manière que la cause s'est produite avant son effet.

Définition 1.7. Cohérence dans le temps : un graphe causal $\mathcal{G}$ pour une série chronologique multivariée est dit cohérent dans le temps si toutes les relations causales restent constantes en terme de direction dans le temps.

Définition 1.8. Stationnarité forte : un processus temporel à valeurs réelles et en temps discret $Z_1, Z_2, ..., Z_t$ est dit stationnaire au sens fort si pour toute fonction f mesurable, $\forall k\, f(Z_1, Z_2, ..., Z_t)$ *et* $f(Z_{1+k}, Z_{2+k}, ..., Z_{t+k})$ ont la même loi.

Définition 1.9. Stationnarité faible : un processus temporel à valeurs réelles et en temps discret $Z_1, Z_2, ..., Z_t$ est stationnaire au sens faible si :
$E[Z_i] = \mu \qquad \forall i = 1...t$
$Var[Z_i] = \sigma^2 \neq \infty \qquad \forall i = 1...t$
$Cov[Z_i, Z_{i-k}] = f(k) = \rho_k \qquad \forall i = 1...t, \quad \forall k = 1...t$

2 La découverte causale

Le domaine de l'apprentissage de structure causale ou de découverte causale à partir de données observationelles vise à reconstruire G_c à partir de D. La plupart des méthodes de découverte causale se divisent en deux groupes : les méthodes basées sur les scores bayésiens d'une part, qui supposent que $p(V)$ a été générée à partir d'un réseau bayésien et visent à trouver le graphe le mieux adapté avec des scores de vraisemblance et, d'autre part, les approches basées sur les contraintes qui essaient de reconstruire le graphe à partir des données avec des itérations de tests statistiques.

2.1 Approches basées sur les scores

Les méthodes basées sur les scores visent à faire correspondre les graphes d'une classe d'équivalence aux données disponibles et évaluer par un score cette correspondance. La formalisation de cette approche a été introduite par Geiger et Heckerman [22, 23] ainsi que Chickering [24]. A partir du jeu de données $\mathcal{D}$ d'un vecteur de variables V, trouvez le graphe $\hat{\mathcal{G}}$ qui maximise un score de vraisemblance $S(\mathcal{D}, \mathcal{G})$:

$$\hat{\mathcal{G}} = \text{argmax}_{\mathcal{G}} S(\mathcal{D}, \mathcal{G}) \tag{2.1}$$

où $\mathcal{G}$ est recherché dans l'espace des DAG.

Nous pouvons avoir des définitions différentes pour la fonction de score. Si la distribution $p(V)$ peut être décrite avec un modèle paramétrique (*e.g.* distributions multinomiales discrètes, relations gaussiennes linéaires), alors on peut définir un ensemble de paramètres $\theta \in \Theta$. La définition bayésienne de $S(\mathcal{D}, \mathcal{G})$ est le log a posteriori avec les croyances a priori $p_{pr}(\mathcal{G})$ et $p_{pr}(\theta)$ sur les DAG et les paramètres respectivement :

$$S(\mathcal{D}, \mathcal{G}) = \log p_{pr}\mathcal{G} + \log p(\mathcal{D}|\mathcal{G}) \tag{2.2}$$

avec $p(\mathcal{D}|\mathcal{G})$ la vraisemblance marginale

$$p(\mathcal{D}|\mathcal{G}) = \int_{\theta \in \Theta} p(\mathcal{D}|\mathcal{G}, \theta) p_{pr}(\theta)$$

Dans cette vue, $\hat{\mathcal{G}}$ qui maximise le score est l'estimateur maximum à postériori.

Une autre façon de définir la fonction de score consiste à utiliser l'estimateur du maximum de vraisemblance $\hat{\theta}$ à partir de N échantillons observés, pour chaque graphe. Nous pouvons ensuite définir la fonction de score à l'aide du critère d'information bayésien (BIC) [25] :

$$S(\mathcal{D}, \mathcal{G}) = \log p(\mathcal{D}|\hat{\theta}, \mathcal{G}) - \frac{d}{2} \log N \tag{2.3}$$

ce qui évite le sur-entraînement en privilégiant les modèles avec moins de paramètres d.

Comme l'espace de recherche croît de façon super-exponentielle avec V, certaines approches utilisent des algorithmes gloutons ou des heuristiques pour limiter l'espace de recherche et fournir une solution adéquate dans un délai raisonnable. Les algorithmes gloutons parcourent l'ensemble des graphes voisins, sélectionnant le meilleur candidat à chaque étape et l'utilisant comme nouveau point de référence. Les voisins sont généralement définis comme tous les DAG qui diffèrent avec au plus une arête manquante ou supplémentaire par rapport au graphe de référence.

L'algorithme Greedy Equivalence Search (GES) est l'une des méthodes basées sur les scores les plus connues pour la découverte causale [24]. Contrairement à l'algorithme PC qui commence par un graphe complet non orienté, GES commence par un graphe vide, sans arête, juste avec les nœuds. Il y a d'abord une phase avant dans laquelle on ajoute des arêtes jusqu'à un maximum local. Ensuite, il y a la phase arrière, où on simplifie le modèle en supprimant des arêtes, le graphe élagué est renvoyé lorsqu'un maximum est atteint.

Les méthodes basées sur les scores souffrent cependant des inconvénients suivants :

(1) La fonction de score nécessite une modélisation et des paramètres simples, qui peuvent détruire les signaux de causalité subtils dans les données réelles.

(2) L'espace de recherche est limité aux DAG ou à leur classe d'équivalence.

(3) Les méthodes ne supportent pas bien l'augmentation de V et ne produisent généralement pas de bons résultats lorsque $V > 50$.

2.2 Approches basées sur les contraintes

Par rapport aux méthodes basées sur les scores, les algorithmes basés sur les contraintes ont une approche plus locale de la reconstruction des graphes. Ils font deux hypothèses : la d-séparation dans $\mathcal{G}_c$ implique une indépendance conditionnelle dans $p(V)$ (la condition de Markov), et toutes les indépendances conditionnelles dans $p(V)$ correspondent à la d-séparation dans $\mathcal{G}_c$ (l'hypothèse de fidélité). Compte tenu des deux hypothèses, les approches basées sur les contraintes sont capables de récupérer jusqu'à la classe d'équivalence de $\mathcal{G}_c$ à partir de l'ensemble des dépendances et des dépendances conditionnelles de $p(V)$.

2.2.1 Principe

Pour illustrer le principe des méthodes basées sur les contraintes, je vais présenter un exemple avec un jeu de données fictif sur le cours des smartphones et des produits les composant.

Cours lithium	Cours batterie	Cours indium	Cours écran	Cours smartphone
0.4812	2.4762	0.5076	5.6766	197.533
0.4321	2.7514	0.5039	6.5099	197.189
0.4181	3.0411	0.49727	7.283	197.361
0.4144	3.3228	0.50283	8.006	196.002
0.4129	3.5801	0.51552	8.6733	193.036
0.4108	3.8007	0.53854	9.2672	188.983
0.4076	3.9767	0.56377	9.7816	185.027
...	...	...	...	...

TABLE 2.1 – Exemple fictif de cours des smartphones, de leurs composants et des matières premières de ces composants.

Le graphe causal attendu est le suivant :

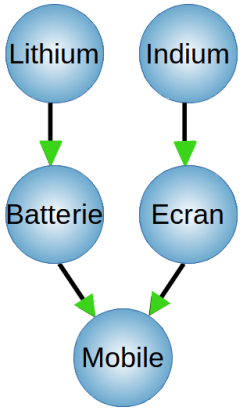


FIGURE 2.1 – **Graphe causal attendu** : le réseau attendu comporte des liens causaux génuines (dont nous sommes sûrs de la cause et de l'effet, représentés par des pointes de flèches vertes) $Lithium \rightarrow Batterie$, $Batterie \rightarrow Mobile$, $Indium \rightarrow Ecran$ et $Ecran \rightarrow Mobile$.

Pour reconstruire ce graphe causal, les méthodes basées sur les contraintes vont procéder en deux étapes : le squelette puis l'orientation du squelette obtenu.

L'étape du squelette part d'un graphe complet dont on va retirer itérativement les arêtes en utilisant un test d'indépendance non conditionnelle entre les deux nœuds de l'arête (X indépendant de Y est noté $X \perp\!\!\!\perp Y$) :

- $X \perp\!\!\!\perp Y \Rightarrow$ l'arête est retirée
- $X \not\perp\!\!\!\perp Y \Rightarrow$ l'arête est conservée

A l'issue de la phase des tests d'indépendance non conditionnelle, une seconde phase de tests d'indépendance conditionnelle va être effectuée. Le principe est de choisir un ensemble de nœuds, les contributeurs $\{Z_i\}$, pour conditionner le test d'indépendance entre les nœuds de chaque arête à évaluer. Les méthodes varient sur la manière de choisir ces contributeurs et selon les hypothèses retenues (comme le fait d'interdire ou d'autoriser la présence de variables latentes).

- $X \perp\!\!\!\perp Y \mid \{Z_i\} \Rightarrow$ l'arête est retirée
- $X \not\perp\!\!\!\perp Y \mid \{Z_i\} \Rightarrow$ l'arête est conservée

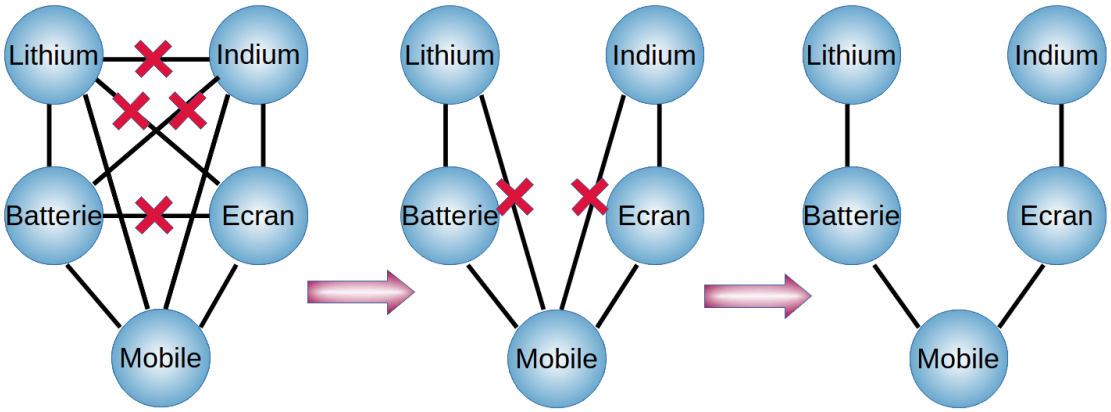


FIGURE 2.2 – **Etape du squelette** : A partir d'un réseau complet, les arêtes sont retirées sur le base de tests d'indépendance non conditionnelle, puis conditionnelle pour parvenir au squelette.

Sur l'exemple des smartphones, pour les arêtes *Lithium* — *Mobile* et *Indium* — *Mobile*, il n'y pas d'indépendance non conditionnelle mais ces arêtes vont pouvoir être supprimées en conditionnant sur, respectivement, *Batterie* et *Ecran*.

A partir du squelette obtenu, la phase d'orientation va consister à détecter les signatures de la causalité dans les données. Pour cela, la méthode va extraire l'ensemble des triplets ouverts comme illustré sur la figure ci-dessous :

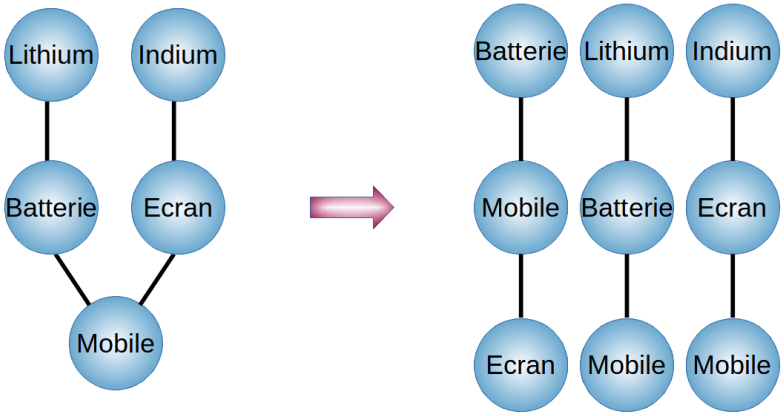


FIGURE 2.3 – **Extraction des triplets ouverts.** Un triplet ouvert est un ensemble de trois nœuds $X, Y, Z \mid X - Z - Y et X \neq Y$.

Ces triplets ouverts vont être évalués pour identifier ceux qui forment une structure en V. Une structure en V est caractérisée par deux conditions : $X \perp\!\!\!\perp Y$ et $X \not\perp\!\!\!\perp Y \mid Z$ et, dans ce cas, les arêtes peuvent être orientées $X \rightarrow Z \leftarrow Y$, c'est la signature de la causalité. Dans le cas général, une structure en V correspond à et $X \perp\!\!\!\perp Y \mid \mathcal{S}$ où $\mathcal{S}$ est l'ensemble de séparation des variables X et Y et $X \not\perp\!\!\!\perp Y \mid \mathcal{S}, Z$ où $Z \notin \mathcal{S}$.

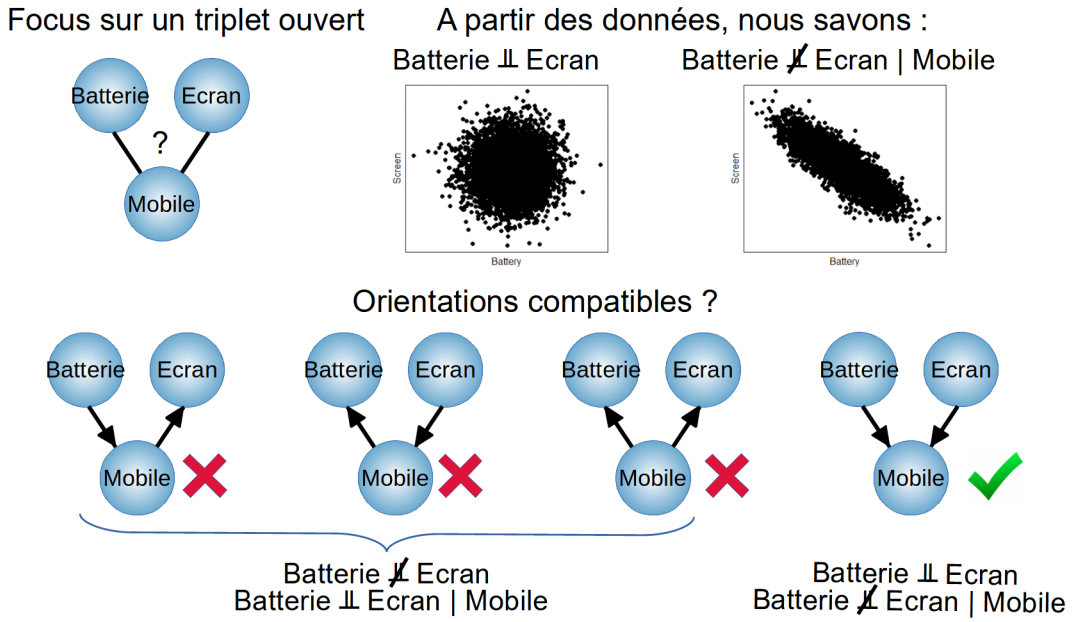


FIGURE 2.4 – **Identification des structures en V à partir des données.** Sur l'exemple des smartphones, on peut logiquement penser que les cours de *Batterie* et *Ecran* doivent être indépendants mais que, dès que l'on fixe le prix du *Mobile*, une dépendance se crée entre *Batterie* et *Ecran* puisqu'un changement de prix de l'un des composants aura tendance à être compensé par l'autre composant afin que le cours du *Mobile* reste inchangé.

A l'inverse, si nous considérons les triplets ouverts $X - Z - Y$ avec les conditions $X \not\perp\!\!\!\perp Y$ et $X \perp\!\!\!\perp Y \mid Z$, aucune orientation ne peut être déduite puisque les orientations $X \leftarrow Z \rightarrow Y$, $X \leftarrow Z \leftarrow Y$ et $X \rightarrow Z \rightarrow Y$ sont toutes possibles pour ces résultats (classe d'équivalence de Markov).

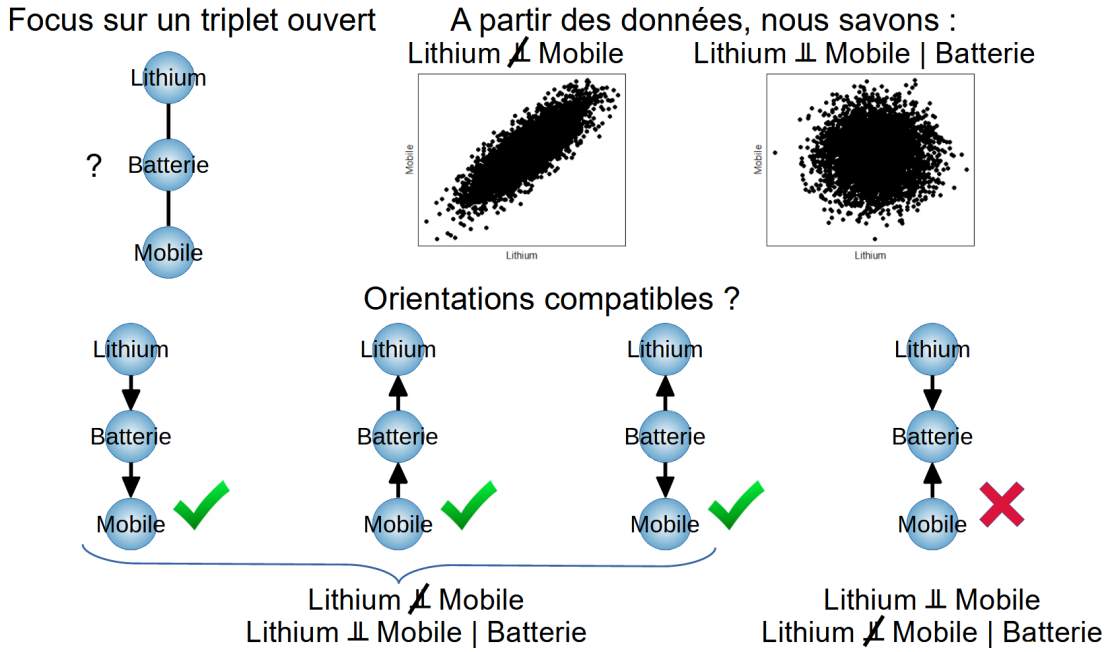


FIGURE 2.5 – **Exemple d'orientation indéterminée.** Sur l'exemple des smartphones, le schéma d'indépendance non conditionnelle et conditionnelle du triplet ouvert *Lithium — Batterie — Mobile* correspond à plusieurs orientations et ne permet pas d'orienter le triplet.

Une fois que l'ensemble des structures en V a été identifié, les orientations sont agrégées et des phases supplémentaires de propagation peuvent éventuellement être appliquées pour parvenir au graphe final. Sur notre exemple des smartphones, le graphe final serait semi-orienté puisqu'aucune signature dans les données ne permet d'orienter les arêtes *Lithium — Batterie* et *Indium — Ecran*.

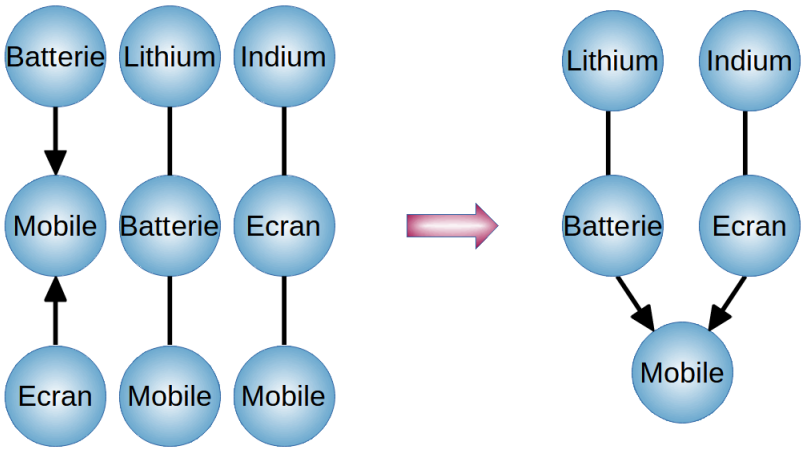


FIGURE 2.6 – **Graphe final**, agrégation des orientations obtenues à partir des triplets ouverts et graphe final semi-orienté

2.2.2 L'algorithme PC

La méthode basée sur les contraintes la plus célèbre pour la découverte causale est l'algorithme PC, nommé d'après Peter Spirtes et Clark Glymour [26, 27]. La version initiale se compose de trois phases, comme indiqué dans l'algorithme 1.

Algorithm 1 Reconstruction causale par l'algorithme PC original

Require: $\mathcal{D}$ the dataset with the V variables

Let $A_{C^{ab}}$ denote the set of vertices adjacent to a or to b in graph C, except for a and b themselves. Let $U_{C^{ab}}$ denote the set of vertices in graph C on (acyclic) undirected paths between a and b, except for a and b themselves (since the algorithm is continually updating C, $A_{C^{ab}}$ and $U_{C^{ab}}$ are constantly changing as the algorithm progresses)

A. Initialization
Form the complete undirected graph C on the vertex set V

B. Squeleton
$n = 0$
repeat
  for each pair of variables (a,b) adjacent in C **do**
    if $A_{C^{ab}} \cap U_{C^{ab}}$ has cardinality greater than or equal to n and a, b are independent conditional on any subsets of $A_{C^{ab}} \cap U_{C^{ab}}$ of cardinality n **then**
      delete $a - b$ from C
    end if
  end for
  $n = n + i$
until for each pair of adjacent vertices a, b, $A_{C^{ab}} \cap U_{C^{ab}}$ is of cardinality less than n
Call the resulting undirected graph F

C. Orientation
for each triple of vertices (a,b,c) such that the pair (a,b) and the pair (b,c) are each adjacent in F but the pair (a,c) are not adjacent in F **do**
  if and only if a and c are dependent one every subset of $A_{F^{ac}} \cap U_{F^{ac}}$ containing b **then**
    Orient $a - b - c$ as $a \rightarrow b \leftarrow c$
  end if
end for

Output all graphs consistent with these orientations

Cette version initiale de PC a maintenant plus de trente ans et a fait l'objet de plusieurs modifications. Pearl et Verma ont notamment décrit comment améliorer l'algorithme en mémorisant les ensembles de séparation lors de l'étape du squelette pour les ré-utiliser lors de l'orientation [26]. D'autres changements ont également été apportés pour parvenir à la version que nous connaissons actuellement comme la possibilité de retourner un CPDAG (completed partially directed acyclic graph) ainsi qu'une étape de propagation des orientations : les orientations sont propagées au reste du graphe selon les règles de Meek [28]. Cette étape ne s'appuyant pas sur les indépendances observées dans les données, il s'agit plutôt d'une convention pour transformer le résultat G_{Inf} en un graphe orienté au maximum dans la classe d'équivalence. En tant que telles, les orientations propagées peuvent être considérées comme des signaux de causalité plus faibles que les structures en V.

Le version de PC que nous connaissons actuellement comprend quatre phases et correspond à l'algorithme 2 décrit ci-dessous :

Algorithm 2 Reconstruction causale par l'algorithme PC

Require: $\mathcal{D}$ the dataset with the V variables

let $Adjacencies(C, A)$ be the set of vertices adjacent to A in directed acyclic graph C (in the algorithm, the graph C is continually updated, so $Adjacencies(C, A)$ is constantly changing as the algorithm progresses)

A. Initialization

form the complete undirected graph C on the vertex set V

B. Squeleton

$n = 0$

repeat

  repeat

    select an ordered pair of variables X and Y that are adjacent in C such that $Adjacencies(C, X)\backslash\{Y\}$ has cardinality greater than or equal to n, and a subset S of $Adjacencies(C, X)\backslash\{Y\}$ of cardinality n

    if X and Y are d-separated given S **then**

      delete edge $X - Y$ from C and record S in $Sepset(X, Y)$ and $Sepset(Y, X)$

    end if

  until all ordered pairs of adjacent variables X and Y such that $Adjacencies(C, X)\backslash\{Y\}$ has cardinality greater than or equal to n and all subsets S of $Adjacencies(C, X)\backslash\{Y\}$ of cardinality n have been tested for d-separation

  $n = n + 1$

until for each ordered pair of adjacent vertices X, Y, $Adjacencies(C, X)\backslash\{Y\}$ is of cardinality less than n

C. Orientation

for each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C **do**

  if and only if Y is not in $Sepset(X, Z)$ **then**

    orient $X - Y - Z$ as $X \rightarrow Y \leftarrow Z$

  end if

end for

D. Propagation

repeat

  if $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B **then**

    orient $B - C$ as $B \rightarrow C$

  end if

  if there is a directed path from A to B, and an edge between A and B **then**

    orient $A - B$ as $A \rightarrow B$

  end if

until no more edges can be oriented

Il a été prouvé que l'algorithme PC était consistant asymptotiquement, renvoyant la classe d'équivalence correcte de G_c si suffisamment d'échantillons sont observés. Cependant, bien qu'il soit théoriquement correct, l'algorithme PC a des limites.

L'une des limitations les plus connues de PC est que les données bruitées/finies peuvent conduire le test d'indépendance à donner un résultat différent du vrai graphe, et cela d'autant plus si ces erreurs se produisent précocement car elles se propagent à d'autres tests et décisions pendant la suite de l'algorithme. En outre, l'ordre dans lequel l'indépendance des variables est testé peut conduire à des résultats différents, ce qui a été corrigé par une version ultérieure appelée PC-Stable [29].

Des modifications ont également été proposées pour améliorer l'étape d'orientation, connues sous la dénomination des règles conservatrice et majoritaire [29, 30]. Avec la règle conservatrice, les structures en V ne sont orientées que lorsque le collisionneur Z n'est dans aucun des ensembles de séparation satisfaisant $X \perp\!\!\!\perp Y|Ui$ alors qu'avec la règle de la majorité, tant que le collisionneur Z est dans Ui moins de 50 % du temps, il est possible d'orienter le collisionneur selon une structure en V.

Une autre limitation de l'algorithme PC est que les ensembles de séparation dans le réseau final ne sont parfois pas cohérents avec la définition de la d-séparation dans les graphes [31]. Par exemple, X et Y ont été rendus indépendants conditionnellement à Z mais Z n'est sur aucun chemin indirect entre X et Y.

2.3 Autre approche

Une approche très différente de la reconstruction des graphes causaux a été proposée avec le modèle linéaire, non gaussien et acyclique (LiNGAM) de Shimizu et al. [32]. Au lieu d'examiner uniquement les dépendances entre variables, dans cette méthode, la relation entre deux variables X et Y est modélisée avec une équation structurelle :

$$Y = bX + \epsilon \tag{2.4}$$

avec b un facteur, ϵ un bruit non gaussien tel que $\epsilon \perp\!\!\!\perp X$. L'apport de Shimizu et al. est la preuve que G_c peut être récupéré à partir de D, dans son intégralité, lorsque tous (ou tous sauf un) les termes d'erreurs sont non gaussiens. L'intuition derrière LiNGAM est que, pour les distributions non gaussiennes, il y a plus d'informations dans la distribution conjointe que dans la matrice de covariance, qui peuvent être détectées par une analyse de composants indépendants. L'hypothèse clé ici est le modèle de bruit additif et indépendant, qui peut être interprété comme le résidu après avoir prédit Y à partir de ses parents. DirectLiNGAM introduit une autre façon de trouver l'ordre causal en effectuant récursivement une régression et un test d'indépendance entre le prédicteur et le résidu [32].

Dans une idée similaire, la méthode Causal Additive Model (CAM) vise à récupérer le DAG sous-jacent en modélisant la distribution $p(V)$ avec des modèles d'équations structurelles additives avec du bruit gaussien et des relations non paramétriques et non linéaires [33].

2.4 La méthode MIIC

La méthode MIIC (Multivariate Information-based Inductive Causation) non temporelle combine les approches basées sur les contraintes avec celles de la théorie de

l'information pour apprendre des graphes causaux plus robustes. Des méthodes basées sur les contraintes, MIIC reprend les étapes classiques, à savoir l'établissement du squelette puis son orientation.

MIIC a été développée sur la base de l'algorithme 3off2 d'Affeldt and Isambert [34, 35], qui utilise l'information multivariée pour reconstruire le squelette et orienter les arêtes. L'information mutuelle corrigée pour les effets de taille finie I′ est utilisée comme proxy pour l'indépendance conditionnelle entre les variables :

$$X \perp\!\!\!\perp Y \;\Leftrightarrow\; \mathrm{I}'(X;Y) \leqslant 0 \text{ et } X \perp\!\!\!\perp Y|Z \;\Leftrightarrow\; \mathrm{I}'(X;Y|Z) \leqslant 0$$

Comme l'algorithme PC, le point de départ est un graphe complet duquel MIIC retire des arêtes pour construire le squelette qui sera ensuite orienté mais il existe une différence cruciale dans la manière dont les ensembles de séparation sont déterminés. Alors que PC boucle sur toutes les combinaisons des voisins de X et Y en augmentant la cardinalité de l'ensemble de séparation $\{U_i\}$ (jusqu'à ce que l'indépendance conditionnelle soit établie ou qu'il n'y ait plus de combinaison possible), MIIC collecte les contributeurs un par un, en utilisant la règle d'enchaînement des informations conditionnelles :

$$I(X;Y|\{U_i\},Z) = I(X;Y) - I(X;Y;U_1) - I(X;Y;U_2|U_1) - \cdots - I(X;Y;Z|\{U_i\})$$

Ceci permet à la fois d'améliorer la rapidité du processus, en supprimant la recherche combinatoire, et le rendre plus robuste aux indépendances fallacieuses en retirant les contributeurs dans l'ordre de leur information. Le choix des contributeurs U_1 puis U_2, U_3, ... s'effectue en effet en sélectionnant à chaque itération le contributeur ayant le score R maximal.

Formellement, le score $R(X,Y;Z|\{U_i\})$ est le minimum entre les deux conditions que Z contribue effectivement à $I(X;Y|\{U_i\})$:

$$R(X,Y;Z|\{U_i\}) = \min\left(P_{\mathrm{nv}}(XYZ|\{U_i\}), P_b(XY|Z,\{U_i\})\right)$$

Où $P_{\mathrm{nv}}(XYZ|\{U_i\})$ est la probabilité que $X - Z - Y$ ne soit pas une structure en V :

$$P_{\mathrm{nv}}(XYZ|\{U_i\}) = \frac{1}{1+e^{-N\mathrm{I}'(X;Y;Z|\{U_i\})}}$$

et $P_b(XY|Z,\{U_i\})$ la probabilité que la base soit $X - Y$

$$P_b(XY|Z,\{U_i\}) = \frac{1}{1+\frac{e^{-N\mathrm{I}'(X;Z|\{U_i\})}}{e^{-N\mathrm{I}'(X;Y|\{U_i\})}}+\frac{e^{-N\mathrm{I}'(Y;Z|\{U_i\})}}{e^{-N\mathrm{I}'(X;Y|\{U_i\})}}}$$

Les règles d'orientation sont également basées sur des mesures de la théorie de l'information et peuvent aussi être exprimées avec des probabilités[35], ce qui contribue à rendre MIIC plus robuste que PC, même en lui appliquant les règles de conservativité ou de majorité.

Comme FCI [27], MIIC est en outre capable de prendre en compte et découvrir des variables latentes [36], la rendant plus apte à analyser des jeux de données de la vie réelle. De plus, MIIC peut évaluer efficacement les jeux de données mixtes, mêlant variables discrètes et continues, ce qui est difficile pour de nombreuses autres méthodes. MIIC possède en outre un ensemble d'options qui peuvent être activées, comme des coupures de confiance, l'inférence de graphes consistants ou encore l'a priori de variable contextuelle.

Mon principal objectif et ma mission de chercheur, pendant ces trois années passées à l'UMR 168 de l'Institut Curie, ont été de comprendre les problématiques de la causalité temporelle puis d'analyser les points à modifier dans la version classique de MIIC pour mettre en œuvre tMIIC, la version temporelle stationnaire, tout en conservant l'ensemble des fonctionnalités existantes.

Algorithm 3 Reconstruction causale par MIIC

Require: $\mathcal{D}$

- Skeleton reconstruction
$\mathcal{G} \leftarrow$ the complete graph on V
for all edges $X - Y \in \mathcal{G}$ **do**
  if $I'(X;Y) \leqslant 0$ **then**
    Delete edge $X - Y$ from $\mathcal{G}$
    Sepset$\{X, Y\} \leftarrow \emptyset$
  else
    Find most contributing node $Z \in \{\text{adj}(X) \cup \text{adj}(Y)\}$ which maximizes $R(X, Y; Z|\emptyset)$
  end if
end for
while There is a link $X - Y$ with $R(X, Y; Z|\{U_i\}) > 1/2$ **do**
  for Top link $X - Y$ with highest rank $R(X, Y; Z|\{U_i\})$ **do**
    Expand contributing set $\{U_i\} \leftarrow \{U_i\} + Z$
    if $I'(X; Y|\{U_i\}) \leqslant 0$ **then**
      Delete edge $X - Y$ from $\mathcal{G}$
      Sepset$\{X, Y\} \leftarrow \{U_i\}$
    else
      Find next most contributing node $Z \in \{\text{adj}(X) \cup \text{adj}(Y)\}$ and compute $R(X, Y; Z|\{U_i\})$
    end if
    Sort the rank list $R(X, Y; Z|\{U_i\})$
  end for
end while

- Skeleton orientation
Sort list of unshielded triples $\mathcal{L}_c = \{(X, Z, Y)_{X \neq Y}\}$ in decreasing order of $|I'(X; Y; Z|\{U_i\})|$
repeat
  Take $(X, Z, Y)_{X \neq Y} \in \mathcal{L}_c$ with highest $|I'(X; Y; Z|\{U_i\})|$ on which orientation rules can be applied
  if $I'(X; Y; Z|\{U_i\}) < 0$ **then**
    if $(X, Z, Y)_{X \neq Y}$ has no diverging orientation, orient as $X \rightarrow Z \leftarrow Y$
  else if $I'(X; Y; Z|\{U_i\}) > 0$ **then**
    if $(X, Z, Y)_{X \neq Y}$ has one converging orientation, propagate orientation as $X \rightarrow Z \rightarrow Y$
  end if
  Update all orientations of $(X, Z, Y)_{X \neq Y} \in \mathcal{L}_c$
until No additional orientation can be obtained

return $\mathcal{G}$

2.5 Contributions à MIIC

Lors de mon arrivée dans l'équipe, j'ai eu l'opportunité de contribuer aux travaux en cours des autres doctorants, notamment sur la fiabilisation des orientations et le projet de MIIC interprétable (iMIIC).

2.5.1 Amélioration de la fiabilité des orientations

Cette contribution à MIIC peut être comparée à la version conservatrice de PC. Elle permet d'améliorer dans MIIC la fiabilité des orientations déduites pour une faible perte de sensibilité en utilisant le principe de supremum d'information mutuelle.

Il est reconnu depuis longtemps [29, 30] que les orientations prédites par les méthodes basées sur les contraintes sont souvent peu fiables, ce qui a largement limité, en pratique, l'utilisation des méthodes basées sur les contraintes pour découvrir des relations causales dans des données d'observation réelles.

Cette incertitude causale provient du grand nombre d'étapes et de conditions que les méthodes basées sur des contraintes, telles que les algorithmes originaux IC [37] ou PC, doivent réaliser avant de pouvoir déduire l'orientation des arêtes. En effet, ils doivent d'abord apprendre un squelette non orienté, en découvrant les indépendances (conditionnelles) entre toutes les paires de variables avant de déduire l'orientation des structures en V et enfin de propager ces orientations à d'autres arêtes non orientées. Cette longue chaîne de décisions informatiques incertaines conduit à l'accumulation d'erreurs qui limitent, au bout du compte, la précision des étapes finales d'orientation et de propagation des méthodes basées sur les contraintes. En conséquence, les orientations des arêtes réduisent considérablement la précision (ou la valeur prédite positive) des graphes causaux inférés par rapport à leur squelette non orienté. De plus, on sait que les méthodes basées sur les contraintes souffrent en général d'une sensibilité ou d'un rappel (c.a.d. un taux de vrais positifs) beaucoup plus faible que la précision [29, 31]. Cela est lié au fait que les ensembles de séparation utilisés pour supprimer les arêtes dans les (premières) étapes des méthodes basées sur les contraintes ne sont souvent pas cohérents avec le squelette final et les graphes orientés (cf section suivante sur la consistance 2.5.2).

Même si MIIC améliorait considérablement la situation en réduisant le déséquilibre entre précision et rappel, pour toutes les tailles d'échantillons [36, 38] et surpassait les méthodes traditionnelles basées sur les contraintes pour déduire des orientations fiables, une perte substantielle de précision subsistait, notamment pour de petits jeux de données, entre les prédictions du squelette et du graphe orienté, ainsi que l'illustre la figure 2.7.

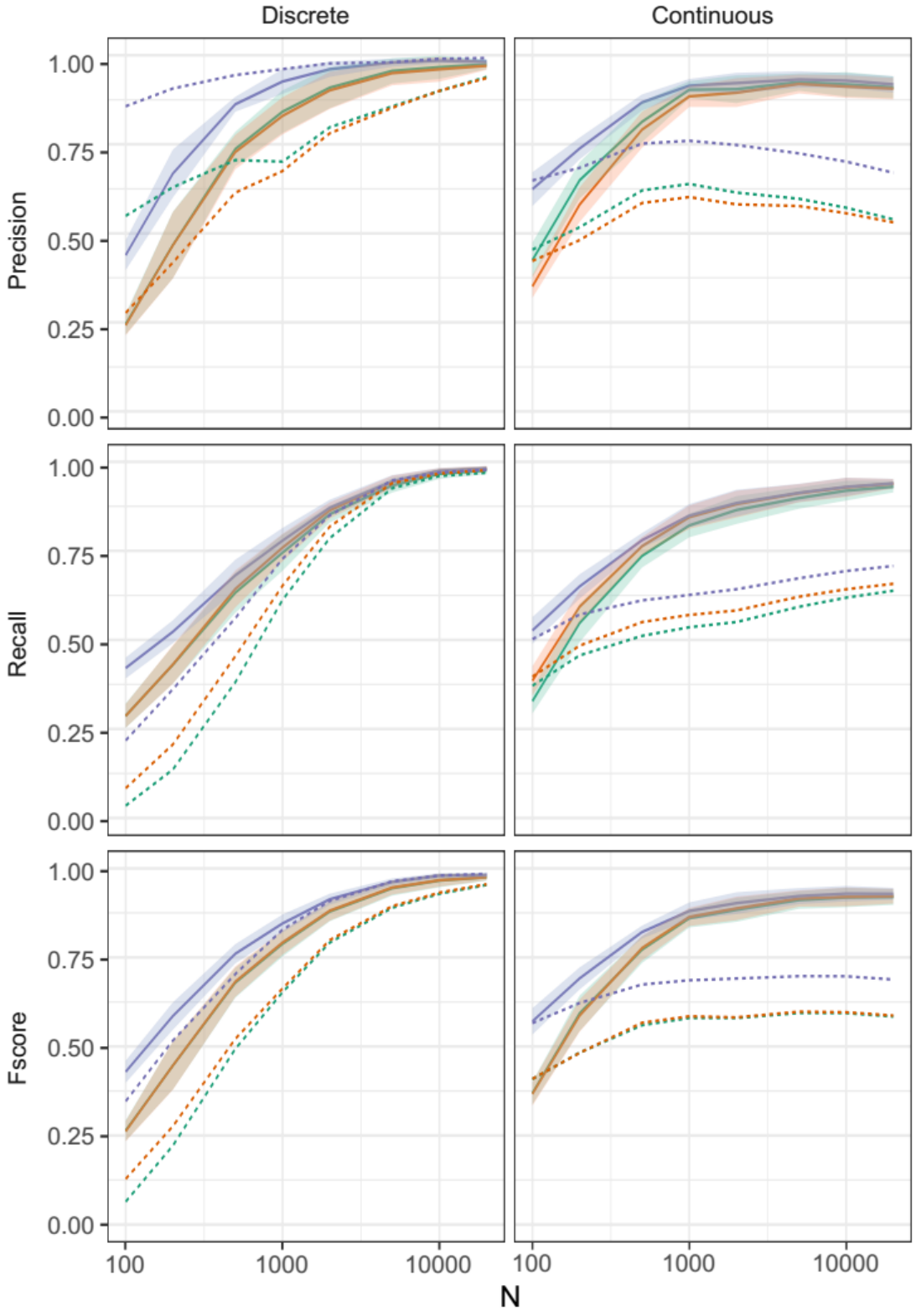


FIGURE 2.7 – **Version initiale de MIIC avec des règles d'orientation autorisant des informations mutuelles (MI) et informations mutuelles conditionnelles (CMI) avec régularisation NML négatives sur des données discrètes (à gauche) et une CMI avec régularisation NML négative sur des données continues (à droite).** Les jeux de données pour les benchmarks sont générés à partir de DAG aléatoires de 100 nœuds avec un degré moyen compris entre 3,8 et 4. Les performances de MIIC sont mesurées en termes de précision, de rappel et de f-score pour le squelette (bleu), le CPDAG (rouge) et le sous-graphe orienté uniquement (vert). Les scores moyens avec PC en utilisant la règle d'orientation majoritaire sont indiqués sous forme de lignes pointillées pour comparaison.

Régles d'orientation initiales

Comme exposé précédemment, MIIC, à partir d'un graphe entièrement connecté, supprime de manière itérative les arêtes superflues, en découvrant des contributions d'informations significatives provenant de chemins indirects en utilisant le principe "3off2" [34, 35]. Cela revient à découvrir progressivement les indépendances conditionnelles les plus conséquentes, c'est-à-dire $I(X;Y|\{A_i\}_n) \simeq 0$ (ou plus formellement $I'(X;Y|\{A_i\}_n) \leq 0$ avec I' définie dans l'eq 2.6), en retirant itérativement les contributions indirectes les plus significatives d'information conditionnelle à 3 points *positives*, $I(X;Y;A_k|\{A_i\}_{k-1}) > 0$, à partir de chaque information (mutuelle) à 2 points, $I(X;Y)$:

$$I(X;Y|\{A_i\}_n) = I(X;Y) - I(X;Y;A_1) - I(X;Y;A_2|A_1) - \cdots - I(X;Y;A_n|\{A_i\}_{n-1}) \tag{2.5}$$

En pratique, l'indépendance (conditionnelle) s'établit en comparant les informations mutuelles (MI) ou informations mutuelles conditionnelles (CMI) à un terme de complexité en maximum de vraisemblance normalisée (NML, Normalized Maximum Likehood) universel, $k_N^{NML}(X;Y|\{A_i\})/N$, calculé sur tous les jeux de données de même taille N et de distributions marginales $p(X,\{A_i\})$ et $p(Y,\{A_i\})$ [34]. Cela peut être vu comme une régularisation NML de la MI et de la CMI pour des jeux de données ayant une taille d'échantillons finie N :

$$I'_N(X;Y|\{A_i\}) = I_N(X;Y|\{A_i\}) - \frac{1}{N} k_N^{NML}(X;Y|\{A_i\}) \tag{2.6}$$

où $k_N^{NML}(X;Y|\{A_i\})$ est calculé de manière itérative en temps linéaire [39, 40] pour un nombre croissant de partitions X et Y, r_x et r_y, en commençant par $k_N^{NML}(X;Y|\{A_i\}) = 0$ pour $r_x = r_y = 1$ [34, 38].

Par conséquent, une indépendance (conditionnelle) est établie pour $I'_N(X;Y|\{A_i\}) \leq 0$, chaque fois que des contributions positives indirectes suffisantes et significatives peuvent être collectées de manière itérative dans l'équation 2.5 pour justifier la suppression de l'arête $X - Y$.

Cela aboutit à un squelette non orienté, que MIIC oriente ensuite (partiellement) en fonction du signe et de l'amplitude des termes d'informations conditionnelles à 3 points régularisées par NML [34, 36], ce qui correspond à la différence entre les termes (C)MI régularisés par NML.

$$I'_N(X;Y;Z|\{A_i\}) = I'_N(X;Y|\{A_i\}) - I'_N(X;Y|\{A_i\},Z) \tag{2.7}$$

En particulier, les termes d'informations conditionnelles à 3 points régularisées par NML négatifs, $I'_N(X;Y;Z|\{A_i\}) < 0$, correspondent à la signature de causalité dans les données observationnelles [34] et conduisent à la prédiction d'une structure en V, $X \rightarrow Z \leftarrow Y$, si $X - Z - Y$ est un triplet ouvert dans le squelette (avec $I'_N(X;Y|\{A_i\}) \leq 0$). En revanche, un terme d'information conditionnelle à 3 points régularisée par NML positif, $I'_N(X;Y;Z|\{A_i\}) > 0$, suggère de propager l'orientation d'une arête précédemment dirigée $X \rightarrow Z - Y$ comme $X \rightarrow Z \rightarrow Y$ (avec $I'_N(X;Y|\{A_i\},Z) \leq 0$), pour correspondre aux hypothèses de la classe de modèle graphique sous-jacent.

Cas des variables discrètes

Cependant, pour les petits jeux de données ou les jeux de données comprenant des variables avec de nombreux niveaux discrets, la complexité NML peut facilement sur-pondérer les termes MI et CMI pour des variables faiblement dépendantes. En conséquence, MIIC a tendance à déduire certaines orientations de structures en V, $X \rightarrow Z \leftarrow Y$, pour lesquelles les deux termes (C)MI régularisés par NML dans l'équation 2.7 sont négatifs : $I'_N(X;Y|\{A_i\}) < I'_N(X;Y|\{A_i\},Z) < 0$, ce qui suggère que Z pourrait en fait être inclus dans un set de séparation de la paire X, Y, en contradiction avec la structure en V inférée $X \rightarrow Z \leftarrow Y$.

Cas des variables continues

Pour les variables continues, MIIC réalise les estimations de MI et CMI via un schéma de discrétisation optimale approximative, basé sur le principe général de supremum de la MI [41] régularisée pour des ensembles de données finis [38]. Cette approche permet de trouver des partitions optimales, P et Q, spécifiant le nombre et les positions des points de coupure de chaque variable continue, X et Y, pour maximiser la MI NML-régularisée entre elles :

$$I'_N(X;Y) = \sup_{P,Q} I'_N([X]_P;[Y]_Q) \tag{2.8}$$

Le terme de régularisation NML, introduit dans $I'_N([X]_P;[Y]_Q)$, est nécessaire pour les jeux de données finis et équivaut à un coût de complexité du modèle, qui finit éventuellement par surpasser le gain d'information en affinant davantage les partitions, lorsqu'il n'y pas suffisamment de données pour étayer un modèle aussi raffiné [38].

Contrairement au cas discret, la perte de précision entre le squelette et le graphe orienté semble différer entre le score CPDAG et le score du sous graphe contenant uniquement les arêtes orientées utilisé pour la comparaison (Fig. 2.8). Cela indique que la précision du sous graphe contenant uniquement les arêtes orientées est légèrement, bien que significativement, meilleure que celle du graphe global partiellement orienté, avec une légère perte concomitante du rappel sur l'orientation, pour de petites tailles d'échantillons (Fig. 2.8). Cette tendance est due à la condition plus stricte d'orientation de la structure en V apportée par les estimations MI NML-régularisées non négatives obtenues par MIIC pour les variables continues. Cependant, le principe de partitionnement optimal ne s'applique qu'à MI [41], et non à CMI, qui doit être estimé par la différence entre les termes optimaux de MI NML-régularisées : $I'_N(X;Y|U) = I'_N(Y;\{X,U\}) - I'_N(Y;U) = I'_N(X;\{Y,U\}) - I'_N(X;U)$ [38].

En conséquence, les estimations approximatives de CMI NML-régularisées entre des variables conditionnellement indépendantes peuvent parfois être négatives et conduire à des orientations de structures en V contredisant l'indépendance conditionnelle, comme évoqué ci-dessus dans le cas discret.

Rectification des MI et CMI régularisées

Le principe général du supremum de MI [41], régularisée dans l'Eq. 2.8 pour les jeux de données finis, est théoriquement valide pour tout type de variable, pas seulement les variables continues. Il pourrait notamment être appliqué à des petits jeux de données avec des variables discrètes ou catégorielles comprenant de nombreux niveaux. Cela entraînerait la fusion des niveaux rares pour mieux estimer les MI et CMI entre des variables discrètes faiblement dépendantes. En fin de compte, les estimations de MI entre variables discrètes indépendantes devraient conduire à la fusion de chaque variable dans un seul groupe, entraînant ainsi un évanouissement complet des estimations de MI NML-régularisées, comme déjà observé pour les variables continues [38]. Par conséquent, la MI NML-régularisée optimale devrait être non négative et, par extension, la CMI NML-régularisée, comme indiqué ici :

Théorème 2.1. *Les optimums de MI et CMI NML-régularisées sont non-négatives*

Preuve. Nous considérons d'abord la MI optimale NML-régularisée, en notant que $I'_N(X;Y) \geq I'_N([X]_1;[Y]_1) \geq 0$ où $[X]_1$ et $[Y]_1$ sont les variables X et Y partitionnées en un unique groupe, ce qui conduit à une disparition de la MI NML-régularisée, car aussi bien la MI que les coûts de complexité NML sont nuls dans ce cas, car $k_N^{NML}(X;Y) = 0$ pour $r_x = r_y = 1$ [34].

Ensuite, la CMI NML-régularisée est définie comme la différence entre les termes de MI NML-régularisées optimaux, soit $I'_N(X;Y|U) = I'_N(Y;\{X,U\}) - I'_N(Y;U) = I'_N(X;\{Y,U\}) - I'_N(X;U)$. Cependant, le partitionnement de X et Y chacun dans un seul regroupement conduit à $I'_N(Y;\{X,U\}) \geq I'_N(Y;\{[X]_1,U\}) = I'_N(Y,U)$ et $I'_N(X;\{Y,U\}) \geq I'_N(X;\{[Y]_1,U\}) = I'_N(X,U)$, ce qui implique que $I'_N(X;Y|U) \geq 0$.

Suite à ces considérations sur la négativité de la (C)MI régularisée par NML avec l'implémentation de l'orientation originale de MIIC, nous avons effectué une modification, basée sur le théorème 2.1 et appelé cette nouvelle version MIIC conservatrice, par analogie avec les règles d'orientation conservatrices des méthodes traditionnelles basées sur les contraintes. [30].

Proposition. MIIC conservatrice rectifie les valeurs négatives des (C)MI NML-régularisées, indiquant une indépendance (conditionnelle), par des valeurs nulles.

Les effets de cette modification sur les benchmarks discrets et continus sont présentés sur la figure 2.8. Bien que la version conservatrice de MIIC n'affecte guère les scores du squelette, il y a clairement un impact sur les scores du CPDAG et des sous-graphes orientés, qui présentent des tendances différentes par rapport à la version initiale de MIIC.

La précision, le rappel et, par conséquent, les f-scores du CPDAG semblent, pour les données discrètes, être légèrement inférieurs avec la version conservatrice (Fig. 2.8) par rapport à la version initiale (Fig. 2.7). Ceci illustre les "meilleurs" scores globaux d'orientation/non-orientation de la version initiale par rapport à l'objectif théorique du CPDAG. En effet, autoriser une MI NML-régularisée négative permet de déduire des structures en V faibles pour de petites tailles d'échantillons. En outre, aucune différence significative n'est observée pour les scores CPDAG sur les données continues, car la version initiale applique déjà une MI NML-régularisée non négative grâce à une optimisation pour les données continues [38], ce qui suggère que l'application de la

non négativité aux CMI NML-régularisées par la version conservatrice de MIIC a peu d'impact sur la fiabilité des scores CPDAG pour les données continues, du moins sur les benchmarks testés ici.

En revanche, la version conservatrice de MIIC améliore considérablement la précision des sous-graphes sur les arêtes orientées uniquement, pour des jeux de données discrets, même pour des tailles d'échantillons relativement petites, Fig. 2.8. Cette forte augmentation de la précision de l'orientation est obtenue au prix d'une perte relativement faible du rappel d'orientation. Par conséquent, MIIC conservatrice améliore considérablement la fiabilité et la sensibilité des orientations prédites pour toutes les tailles d'échantillons, par rapport aux méthodes traditionnelles basées sur des contraintes avec des règles d'orientation conservatrices, Fig. 2.8. Par exemple, MIIC conservative atteint déjà près de 90 % de précision d'orientation avec 25 % de rappel en orientation pour $N \simeq 250$, contre environ 80 % de précision d'orientation avec seulement 5 % de rappel en orientation pour PC conservatrice. Lorsque PC conservatrice atteint une précision d'orientation de 90 % avec un rappel en orientation de 25 % pour $N \simeq 700$, MIIC conservatrice atteint une précision d'orientation de près de 100 % avec un rappel en orientation de 50 %, Fig. 2.8. De plus, alors que la version précédente de MIIC atteint un rappel en orientation nettement meilleur de 65 % pour le $N \simeq 700$, Fig. 2.7, sa précision en orientation chute simultanément à environ 75 %, ce qui a clairement un impact sur sa fiabilité pour la découverte causale.

Sur des données continues, MIIC conservatrice permet également d'obtenir une augmentation importante de la précision de l'orientation, qui devient comparable à la précision du squelette, même pour de petits jeux de données, et est clairement bien meilleure que les scores obtenus avec les méthodes traditionnelles basées sur des contraintes pour de grands ensembles de données, Fig. 2.8. Par exemple, MIIC conservatrice atteint une précision d'orientation de près de 75 % avec un rappel d'orientation de 50 % pour $N \simeq 200$, contre environ 70 % de précision d'orientation avec un rappel en orientation de 35 % pour PC conservatrice. Au moment où PC conservatrice atteint une précision d'orientation de 75 % avec un rappel d'orientation de 45 % pour $N \simeq 1\ 000$, MIIC conservatrice atteint plus de 90 % de précision d'orientation avec 80 % de rappel en orientation, fig. 2.8.

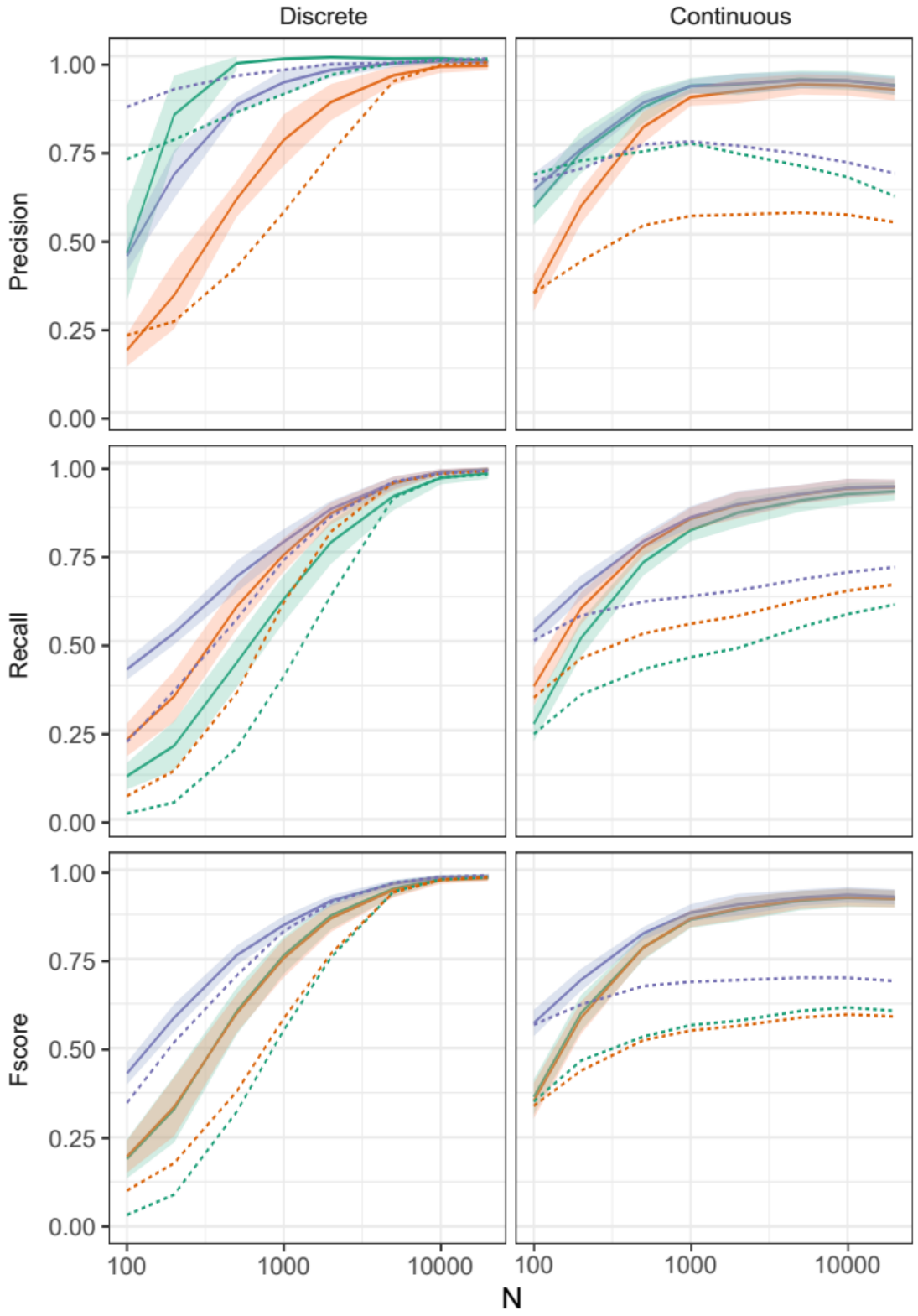


FIGURE 2.8 – **MIIC conservatrice** avec de nouvelles règles d'orientation appliquant des MI et CMI NML-régularisées non négatifs sur des données discrètes (à gauche) ainsi que des données continues (à droite). Les performances d'apprentissage de la structure par MIIC conservatrice sont mesurées en termes de précision, de rappel et de f-scores pour le squelette (bleu), le CPDAG (rouge) et le sous-graphe orienté uniquement (vert). Les scores moyens de PC avec les règles d'orientation conservatrices sont affichés sous forme de lignes pointillées pour comparaison.

2.5.2 MIIC interprétable

Le projet de MIIC interprétable (iMIIC) regroupe un ensemble d'améliorations à la méthode MIIC. Les améliorations incluses dans iMIIC concernaient l'orientation des arêtes, la distinction entre causes "authentiques" et "putatives", la prise en charge des variables contextuelles, la consistance et la capacité à traiter des volumes de données très importants.

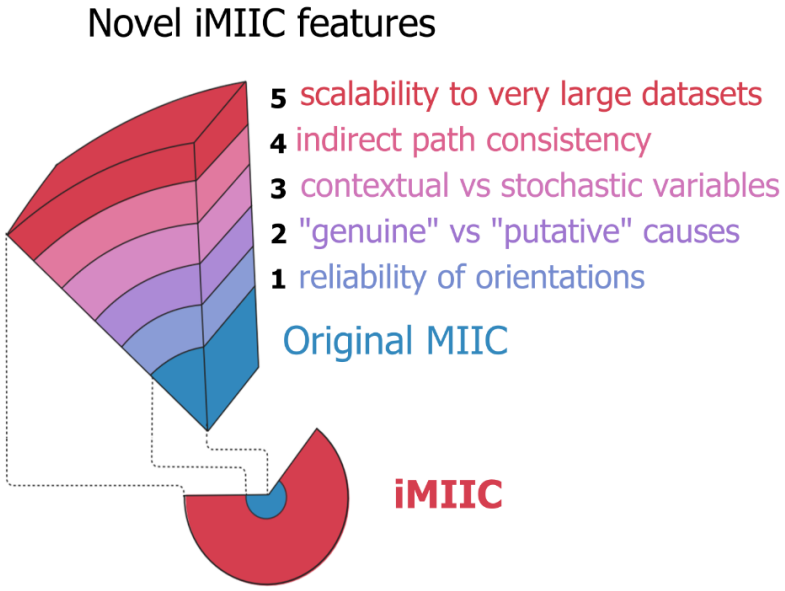


FIGURE 2.9 – MIIC interprétable (iMIIC).

Je ne détaillerai pas la capacité à traiter des volumes de données très importants car il s'agit essentiellement d'une refonte profonde et d'une optimisation du code existant. La partie sur la fiabilité des orientations, quant à elle, a été abordée dans la section précédente sur le principe du supremum d'information mutuelle.

Causes "authentiques" et "putatives"

Dans les algorithmes basés sur des contraintes pour la découverte causale, il est courant d'observer une propagation des orientations. Si nous avons un squelette de graphe avec deux triples ouverts $X - Z - Y$ et $X - Z - T$, où $X - Z - Y$ se révèle être une structure en V ($X \to Z \leftarrow Y$) mais pas $X - Z - T$, alors des algorithmes tels que PC orienteront la non structure en V comme $X \to Z \to T$, car c'est la seule orientation compatible avec un DAG.

iMIIC apporte l'idée de probabilités d'orientations en attribuant une probabilité à chaque extrémité d'arête, ce qui nous permet de distinguer un facteur de confusion latent non mesuré provoquant X et Y (représenté visuellement par $X \leftrightarrow Y$), une arête avec une probabilité suggérant une pointe de flèche sur une seule des extrémités, appelée arête putative ($\to$ ou $\leftarrow$), et une arête pour laquelle il existe des probabilités suggérant une pointe de flèche pour l'une des extrémités et une queue de flèche pour l'autre extrémité, appelée authentique arête causale ($\Rightarrow$ ou $\Leftarrow$).

Pour décider de l'orientation de chaque extrémité, trois règles de base peuvent être considérées :

1. Si la probabilité p est de 0,5, l'orientation de cette extrémité est indéterminée.
2. Si $p > 0,5$, l'orientation de cette extrémité est probablement une pointe de flèche.
3. Si $p < 0,5$, l'orientation de cette extrémité est probablement une queue de flèche.

Cependant, iMIIC fournit un paramètre appelé coupure de probabilité d'orientation $p* \geq 0,5$ qui filtre les probabilités de toutes les extrémités. Au lieu des trois règles ci-dessus comparant toujours à 0,5, pour qu'une extrémité soit considérée comme une pointe de flèche, elle doit avoir une probabilité $p > p*$ et, pour être une queue de flèche $p < 1 - p*$, sinon $(1 - p* < p < p*)$ est indéterminé.

Dans cet esprit, la différence entre une arête causale putative ($\rightarrow$ ou $\leftarrow$) et une arête causale authentique ($\Rightarrow$ ou $\Leftarrow$) est que, dans le cas putatif, nous avons $p > p*$ pour une extrémité (pointe de flèche) mais $1 - p* < p < p*$ pour l'autre extrémité. La condition n'est donc remplie que pour une extrémité, qui a une pointe de flèche. Le cas de l'arête causale authentique, qui apporte une certaine certitude que la queue est bien une queue et que la tête est bien une tête, nous donne une confiance raisonnable qu'il s'agit d'une véritable arête causale.

Variables contextuelles

Cette introduction d'un cadre probabiliste distinguant les orientations de la tête de flèche par rapport à la queue mis en œuvre dans iMIIC permet d'inclure des connaissances préalables sur certaines orientations de la tête ou de la queue. Dans iMIIC, nous avons utilisé cette possibilité pour définir la notion de variable contextuelle et, dans le chapitre 7, je décrirai comment j'ai mis à profit cette fonctionnalité pour ajouter l'a priori de conséquence.

Inclure quelques variables contextuelles dans les modèles graphiques peut permettre de préciser, selon la nature de l'ensemble de données, un paramètre de contrôle ou des conditions expérimentales ou encore de caractériser le profil personnel de patient (par exemple sexe, année de naissance).

Contrairement à la plupart des autres variables du jeu de données, ces variables contextuelles ne varient pas de manière stochastique et doivent, par hypothèse, avoir toutes leurs arêtes sans pointe de flèche entrante, soit $p_{contextual} = 0$. Cela exprime notre connaissance préalable que les variables contextuelles ne peuvent pas être la conséquence d'autres variables, quelles soient observées ou non, dans le jeu de données.

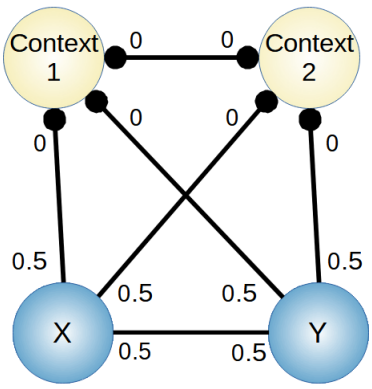


FIGURE 2.10 – Principe de l'ajout dans MIIC de l'a priori de variable contextuelle.

Consistance

Lorsque l'algorithme retourne le graphe final, il est possible que ce graphe soit inconsistant avec les indépendances conditionnelles utilisées pour retirer des arêtes au cours de l'algorithme, comme cela a été observé sur PC [31].

Si l'on imagine qu'à un moment donné, une arête entre X et Y a été supprimée en raison d'un voisin commun Z ($X - Z - Y$), ou d'une variable dans un chemin entre X et Y mais que, dans le graphe final, ce chemin n'existe plus, l'ensemble de séparation utilisé est incompatible avec le squelette observé dans le graphe final.

Au niveau de l'orientation, il est également possible qu'un nœud, qui se trouvait sur le chemin entre X et Y soit toujours dans le chemin dans le graphe final et soit désormais un descendant commun de X et Y, un collisionneur par exemple. Cet ensemble de séparation est également incohérent avec le graphe orienté final, car un descendant commun ne doit pas être utilisé dans l'ensemble de séparation.

Ces deux types d'inconsistance sont illustrés dans les figures 2.11 et 2.12.

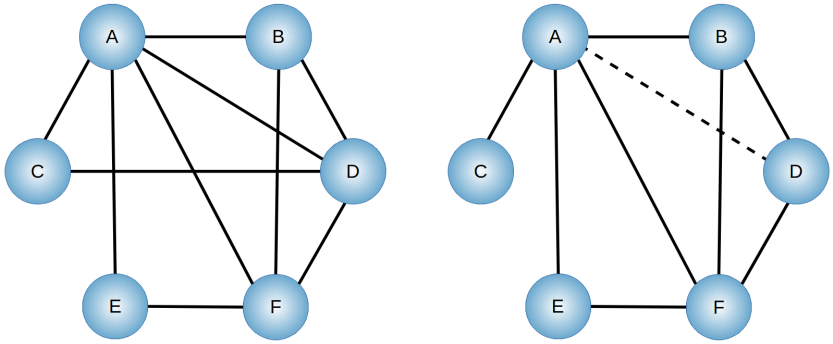


FIGURE 2.11 – **Inconsistance au niveau du squelette.** Le graphe de gauche illustre qu'à un moment antérieur de la reconstruction, il y avait un chemin entre A et D via C avec $A \perp\!\!\!\perp D \mid \{C, F\}$. Cependant, sur le squelette du graphe final à droite, il est possible de voir que l'arête entre C et D a été supprimée ultérieurement, et que C n'est donc plus dans un chemin entre A et D, ce qui rend l'ensemble de séparation C, F incohérent avec le graphe final.

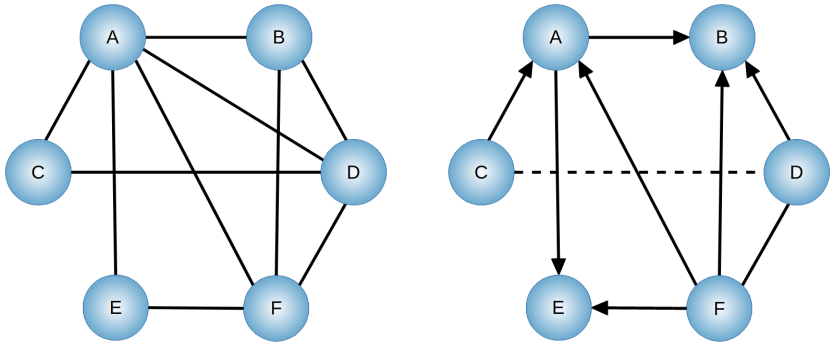


FIGURE 2.12 – **Inconsistance au niveau des orientations.** A un certain point de la découverte causale, l'arête entre C et D a été supprimée car $C \perp\!\!\!\perp D \mid B$. Cependant, dans le graphe final de droite, on voit que B n'était pas un médiateur, mais un collisionneur, un descendant commun de C et D et il ne devrait donc pas être retenu pour l'ensemble de séparation. L'ensemble de séparation B est donc incohérent avec l'orientation du graphe final.

Le contrôle de cohérence de l'ensemble de séparation qui a été implémenté dans iMIIC fonctionne de manière itérative. À la fin de chaque itération, les ensembles de séparation sont vérifiés comme décrit dans les figures 2.11 et 2.12. A l'itération suivante, lors du choix des ensembles de séparation, le graphe obtenu à l'itération précédente est pris en compte.

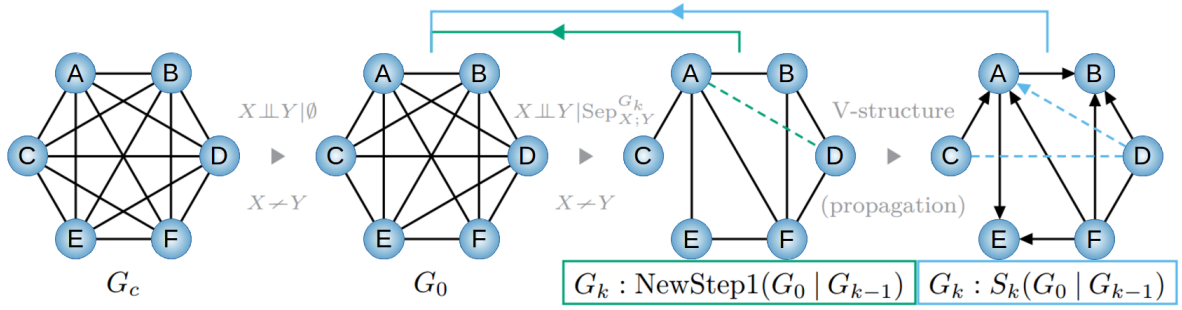


FIGURE 2.13 – **Approche itérative pour apprendre des graphes causaux dont les ensembles de séparation sont cohérents avec le squelette et/ou les orientations.** Les arêtes en pointillés montrent la différence entre deux itérations successives.

Idéalement, l'objectif serait de trouver successivement deux fois le même graphe, mais il n'y a aucune garantie que cela se produira de manière certaine. La règle pour mettre fin à l'itération est donc d'attendre d'obtenir un graphe identique à l'un des graphes précédents, ce qui produit une série de graphes $G_{k-n}, G_{k-n+1}, ..., G_k$ tel que $G_{k-n} = G_k$. L'ensemble $G_{k-n}, G_{k-n+1}, ..., G_k$ est alors appelé un cycle cohérent.

Le graphe consistant final retourné par iMIIC est l'union des graphes du cycle, dont la cohérence est garantie. Cependant, prendre l'union d'un tel cycle pouvant potentiellement ajouter de nombreuses arêtes non orientées et supprimer de nombreuses pointes de flèches, ce qui réduit le caractère informatif du graphe final, nous avons également introduit la notion de graphe consensus, associé à un seuil α_c. Au prix d'une perte potentielle de cohérence des ensembles de séparation, le graphe consensus peut fournir plus d'informations que le graphe consistant. Nous pouvons noter qu'un graphe consensus avec $\alpha_c = 1$ est identique au graphe consistant.

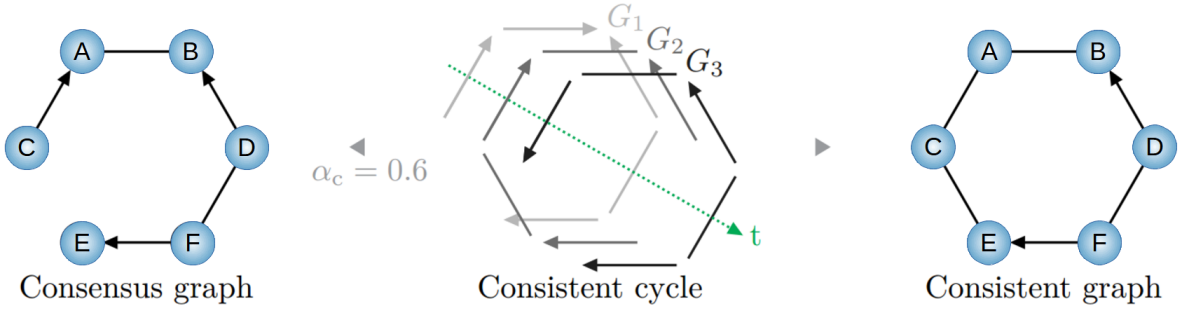


FIGURE 2.14 – **Graphes consistant et consensus.** Au milieu, le cycle consistant de taille 3. A droite, le graphe consistant en terme d'orientation. A gauche, graphe consensus avec $\alpha_c = 0{,}6$.

Benchmark

De nombreux benchmarks, notamment relatifs aux orientations et aux temps d'exécution ont été réalisés pour évaluer les performances de iMIIC et sont disponibles en annexes dans la publication correspondante.

Parmi ces benchmarks, celui présenté en figure 2.15 permet une vue d'ensemble des performances de iMIIC et montre que iMIIC dépasse largement d'autres méthodes de référence.

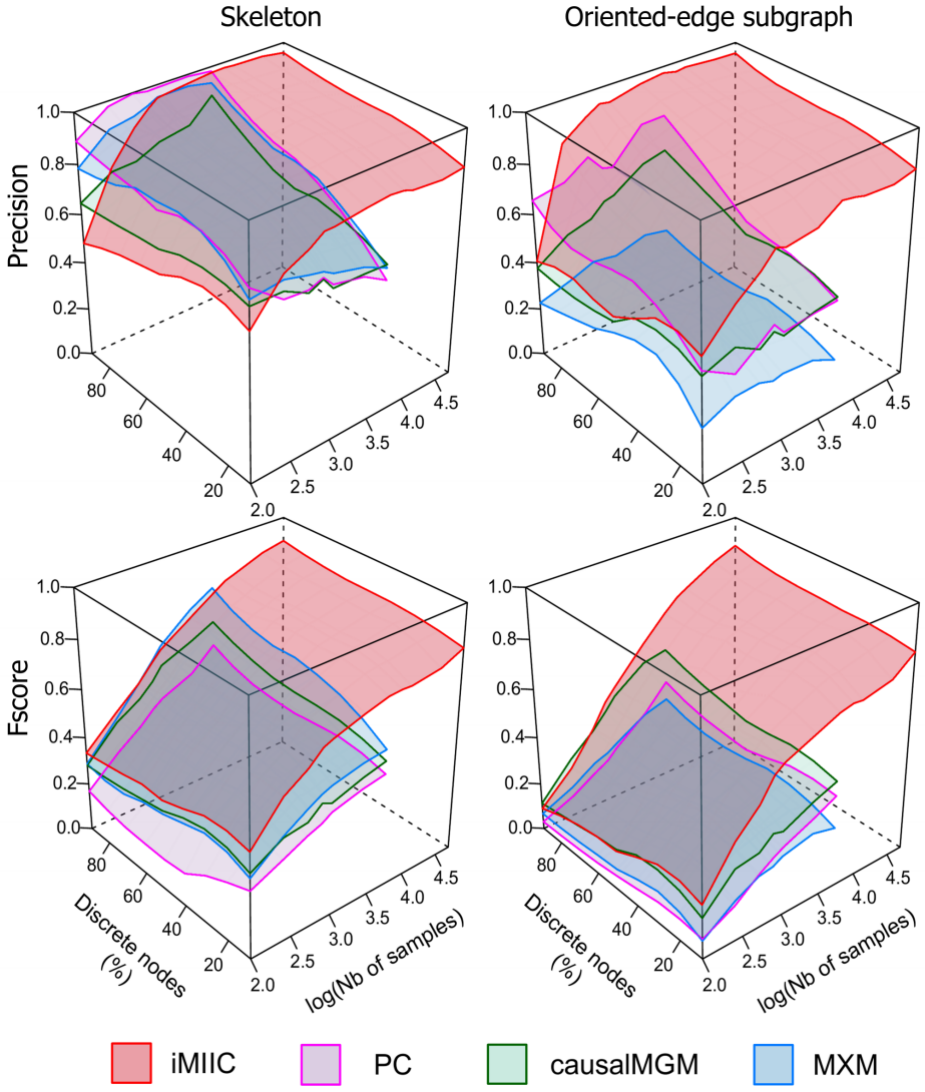


FIGURE 2.15 – **Benchmarks sur des données synthétiques similaires à l'application sur la base de données SEER** avec différentes proportions de variables discrètes et un nombre croissant d'échantillons.

Application

iMIIC a été appliqué sur un vaste ensemble de données concernant le cancer du sein provenant du programme "Surveillance, Epidemiology, and End Results" (SEER) [42] du National Cancer Institute, qui collecte des données à propos du cancer sur les diagnostics, le traitement et la survie des patients environ 35 % de la population des Etats-Unis.

Au sein de la base de données SEER qui comprend plus de 11 millions d'entrées, nous avons extrait les données relatives au cancer du sein pour la période 2010-2016, ce qui représente 407 791 dossiers médicaux pour 396 179 patientes distinctes en raison de tumeurs primitives multiples du sein chez certaines patientes. Cinquante et une variables cliniques, socio-économiques et de dénouement ont été sélectionnées pour leur pertinence vis à vis du cancer du sein, leur faible redondance ainsi que leur niveau d'informations manquantes contenu.

Le réseau de cancer du sein inféré par iMIIC (Fig. 2.16), fournit un modèle graphique interprétable comprenant 280 arêtes, pour lesquelles la plupart des relations de cause à effet sont connues ou peuvent être exclues sur la base de connaissances générales ou expertes ainsi que de la pratique clinique. En terme de temps d'exécution, iMIIC a inféré ce graphe causal en 12 heures et 46 minutes pour les 51 variables et 396 179 échantillons.

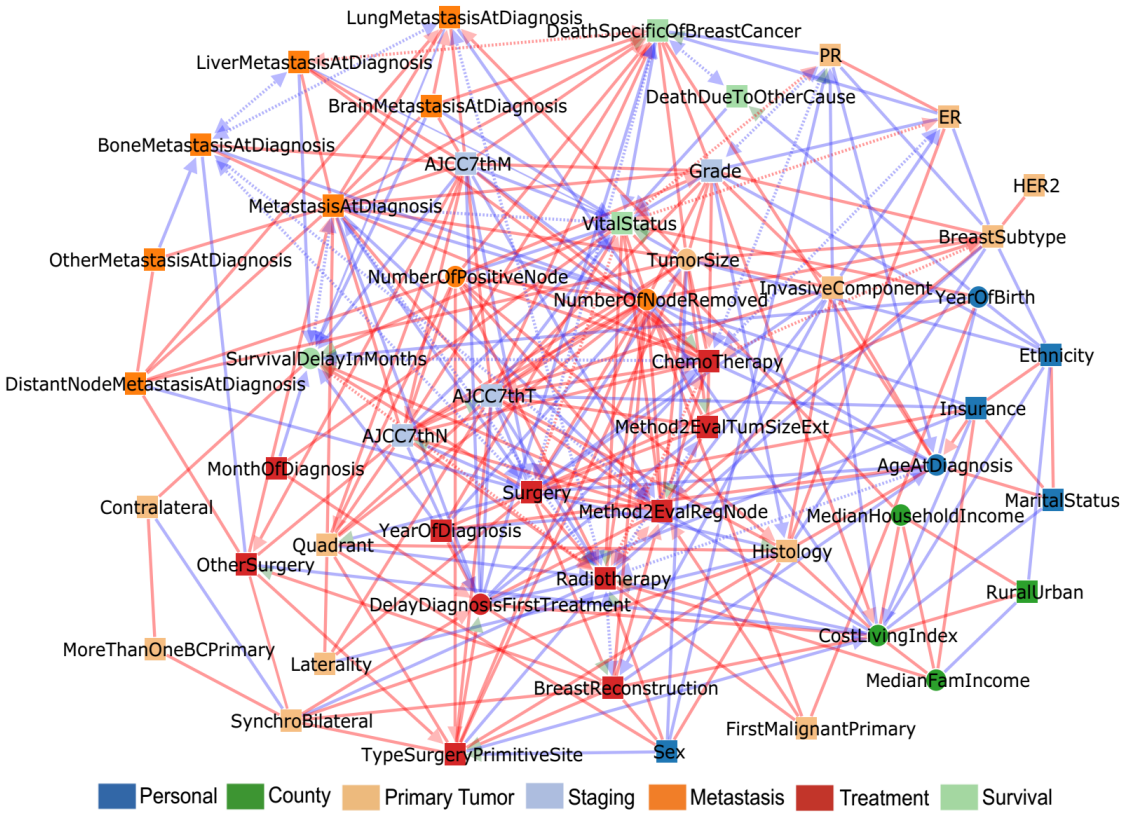


FIGURE 2.16 – **Réseau de 51 nœuds inféré par iMIIC à partir de l'ensemble de données SEER incluant 396 179 patientes atteintes d'un cancer du sein diagnostiquées entre 2010 et 2016.** Ce réseau consistant en terme de squelette contient 280 arêtes et comprend 2 variables contextuelles (le sexe et l'année de naissance).

Globalement, l'évaluation faite avec des experts du domaine indique qu'environ 90 % des effets causaux authentiques ou putatifs déduits par iMIIC sont corrects et que 8 % supplémentaires des relations causales sont plausibles sur la base des connaissances cliniques et épidémiologiques. Par conséquent, les règles d'orientation améliorées de iMIIC conduisent à seulement 2 % d'arêtes causales erronées, contre environ 15 % lorsque les règles d'orientation précédentes de MIIC étaient appliquées pour analyser la même cohorte SEER d'environ 400 000 patientes, ce qui atteste des améliorations apportées par iMIIC.

De nombreux résultats, parfois contre-intuitifs, ont pu être identifiés par les réseaux causaux reconstruits par iMIIC. Sans reprendre l'ensemble des résultats qui sont décrits dans la publication disponible en annexes, il est possible de citer le lien causal de *VitalStatus* vers *Radiotherapy* qui est contre-intuitif de prime abord, puisque nous nous attendons que ce soit la radiothérapie qui améliore les chances de survie. Cependant une relation causale à court terme entre *VitalStatus* et *Radiotherapy* est cohérente avec la baisse rapide de la distribution des délais de survie au cours des 3 à 6 premiers mois en l'absence de radiothérapie (Fig 2.17b). En global, 51 % des patientes vivantes ont subi une radiothérapie contre seulement 27 % des patientes décédées (Fig 2.17a). Ceci suggère qu'un décès prématuré au cours des premiers mois suivant le diagnostic pourrait empêcher la radiothérapie chez certains patientes qui auraient pu autrement recevoir ce traitement si elles avaient vécu plus longtemps. Cet effet causal à court terme correspond à la plage typique des délais de radiothérapie après le diagnostic, selon qu'il est réalisé en deuxième traitement après une intervention chirurgicale ou en troisième traitement après une intervention chirurgicale et une chimiothérapie. Dans l'ensemble, cet effet causal à court terme des décès prématurés sur la radiothérapie l'emporte sur l'effet bénéfique, causalement inversé, de la radiothérapie sur la survie à long terme des patientes.

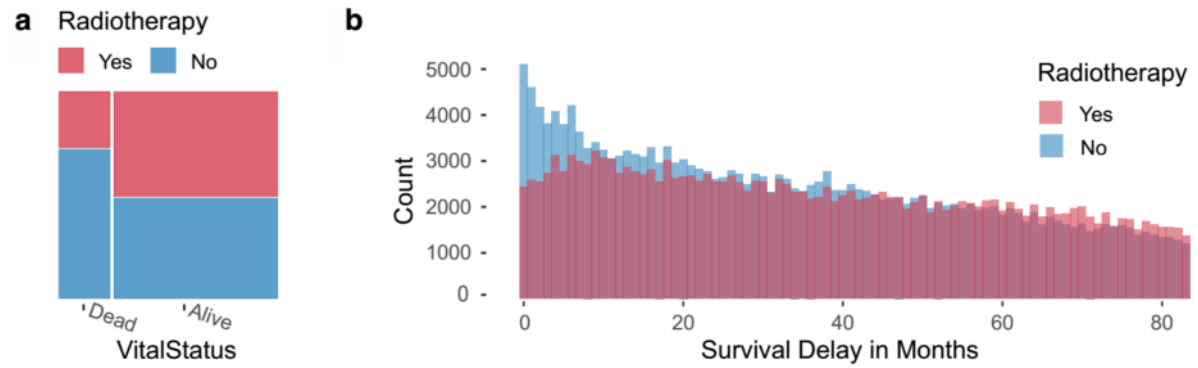


FIGURE 2.17 – **a**, Statut vital avec et sans radiothérapie. **b**, distribution des délais de survie avec et sans radiothérapie.

iMIIC explique également des corrélations connues telles celles entre assurance ou statut marital avec les chances de survie, où l'assurance ou le statut marital ne peuvent être des causses directes des chances de survie. Les réseaux consistants produits par iMIIC permettent ici d'expliquer ces corrélations par des effets indirects, avec pour *Insurance* : *Surgery* (50 %), *ChemoTherapy* (14 %), *MaritalStatus* (20 %), *Radiotherapy* (9 %), *Breast reconstruction* (7 %) et, pour *MaritalStatus* : *Surgery* (58 %), *Year of birth* (40 %), et *Ethnicity* (2 %).

3.1 Panorama des méthodes

De nombreuses méthodes ont été développées pour l'analyse causale des séries temporelles [19, 43]. Elles peuvent être distinguées en deux grandes classes : l'analyse de séries multivariées (MTS) ou l'analyse de séquences d'évènements comme illustré sur la figure ci-dessous :

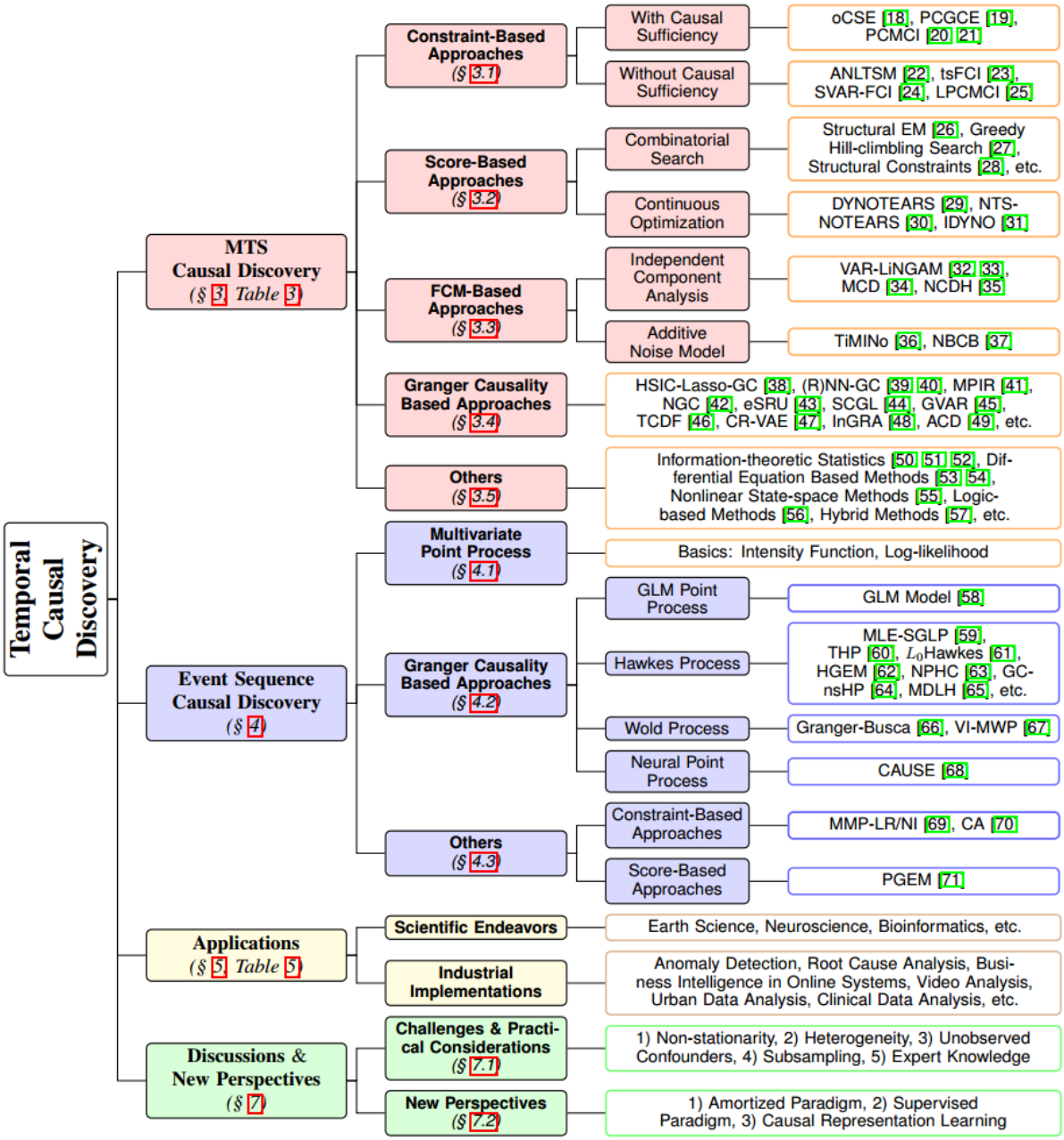


FIGURE 3.1 – **Panorama de la découverte causale temporelle.** (extrait de "Causal Discovery from Temporal Data : An Overview and New Perspectives", C. Gong et al, 2023)

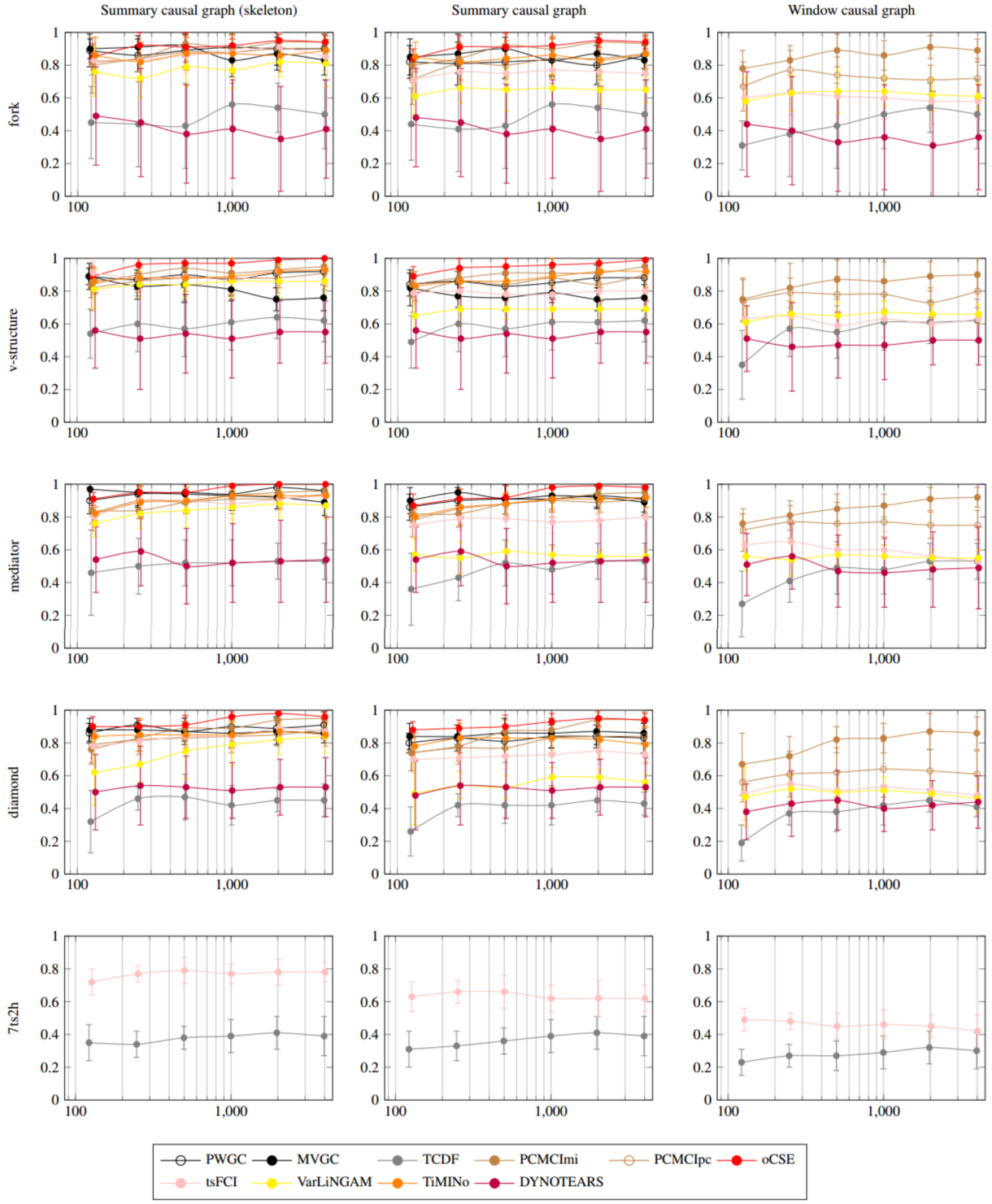


FIGURE 3.2 – **Performance de différentes méthodes de découverte causale temporelle sur 5 jeux de données artificielles.** Les résultats sont calculés sur 10 exécutions pour lesquels sont rapportés la moyenne (± l'écart type) du f-score non orienté pour le graphe sommaire (colonne de gauche), du f-score orienté pour le graphe sommaire (colonne du milieu) et le f-score orienté du graphe fenêtré (colonne de droite). Les résultats sont calculés pour différentes longueurs de séries chronologiques : 125, 250, 500, 1 000, 2 000 et 4 000 pas de temps, une échelle logarithmique est utilisée pour l'axe des x. (extrait de "Survey and Evaluation of Causal Discovery Methods for Time Series", C. K. Assaad et al, 2022)

Dans cette thèse, je développerai plus particulièrement autour de la causalité de Granger, qui est historiquement, une contribution fondamentale dans le domaine de la découverte causale temporelle, ainsi que l'entropie de transfert qui étend la causalité de Granger et qui utilise l'information mutuelle. Enfin, dans notre cas, l'application envisagée en adaptant la méthode MIIC en une version temporelle étant l'analyse de séries multivariées avec un graphe fenêtré et MIIC étant une méthode basée sur les contraintes, nous nous sommes plus spécifiquement intéressés aux méthodes PC, qui est la référence classique et PCMCI qui obtient de bons résultats sur un graphe fenêtré comme illustré sur la figure 3.2.

3.2 L'apport attendu du temps

Dans le cas d'analyse de séries multivariées, si nous reprenons l'exemple précédent sur les smartphones mais en supposant maintenant que nous ajoutons le temps comme information supplémentaire dans le jeu de données, par exemple avec le jour :

Jour	Cours lithium	Cours batterie	Cours indium	Cours écran	Cours smartphone
1	0.4812	2.4762	0.5076	5.6766	197.533
2	0.4321	2.7514	0.5039	6.5099	197.189
3	0.4181	3.0411	0.49727	7.283	197.361
4	0.4144	3.3228	0.50283	8.006	196.002
5	0.4129	3.5801	0.51552	8.6733	193.036
6	0.4108	3.8007	0.53854	9.2672	188.983
...	...	...	...	...	...

TABLE 3.1 – Exemple fictif de cours des smartphones, de leurs composants et des matières premières de ces composants avec l'information de temps

En supposant que les effets sont décalés dans le temps, ce qui vraisemblable dans notre exemple, le graphe obtenu pourrait désormais être totalement orienté en utilisant la temporalité. De plus, nous pourrions même imaginer de révéler des interactions plus complexes comme des boucles de rétro-actions :

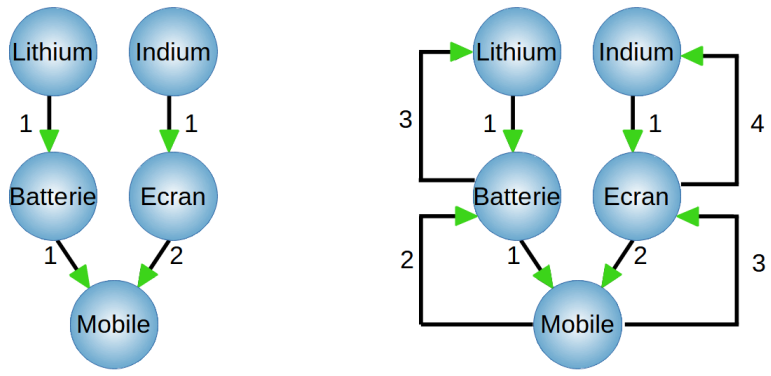


FIGURE 3.3 – **Apport du temps.** En utilisant l'information du temps, le graphe reconstruit peut être complètement orienté, voire révéler de nouvelles dynamiques

3.3 Problématique de base

La problématique de base de la découverte causale temporelle sur des séries multivariées, comme celle de la découverte causale générale, est de reconstituer le graphe causal uniquement à partir de données observationnelles, tout en ajoutant le temps comme contrainte ou information complémentaire.

Par exemple, si nous générons des séries temporelles avec ce modèle :

$$X^1_t = 0.5 * X^1_{t-1} - 0.2 * X^2_{t-1} + \eta_1$$
$$X^2_t = 0.55 * X^2_{t-1} + 0.3 * X^4_{t-3} + \eta_2$$
$$X^3_t = 0.45 * X^3_{t-1} - 0.3 * X^2_{t-2} + 0.2 * X^4_{t-1} + \eta_3$$
$$X^4_t = 0.40 * X^4_{t-1} + \eta_4$$

L'objectif est, uniquement à partir des données, de reconstruire le graphe temporel causal correspondant. Sur cet exemple, la dynamique étant stationnaire, le graphe à inférer peut être infini, car il est invariant par translation dans le temps.

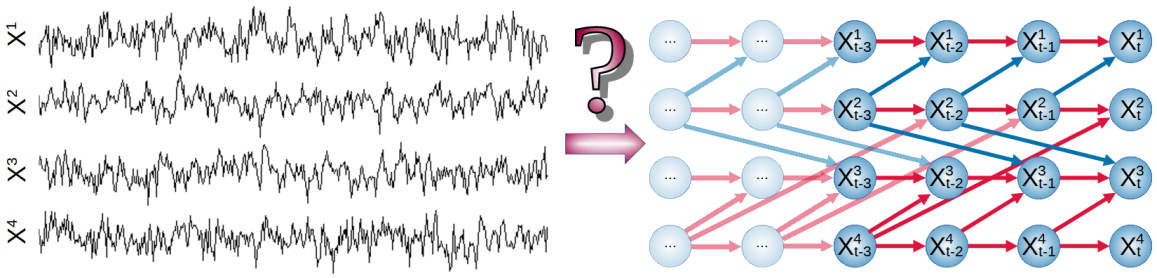


FIGURE 3.4 – **Problématique**, à partir d'observations de l'évolution de variables au cours du temps, comment déterminer le graphe causal temporel correspondant.

3.4 La causalité de Granger

Granger [44] est certainement le nom le plus connu dans le domaine de la causalité temporelle avec ses travaux sur les séries chronologiques qui ont fondamentalement changé la façon dont les économistes analysent les données financières et macroéconomiques et pour lesquels il a reçu le Prix Nobel de sciences économiques.

L'idée de base de la causalité au sens de Granger est qu'une série temporelle X causerait une autre série Y lorsque la connaissance du passé de X modifierait la prévision de Y fondée uniquement sur le passé de Y :

Dans sa version la plus simple, la causalité standard de Granger par paire repose sur l'hypothèse d'un système linéaire stationnaire et, pour déterminer si X Granger-cause Y, il convient d'utiliser les modèles auto-régressifs suivants :

Modèle restreint : $Y_t = a_0 + \sum_{i=1}^{\tau} a_i Y_{t-i} + \epsilon_t$
Modèle complet : $Y_t = a_0 + \sum_{i=1}^{\tau} a_i Y_{t-i} + \sum_{i=1}^{\tau} b_i X_{t-i} + \epsilon_t$

où ϵ_t sont des termes d'erreurs indépendants distribués selon une loi normale, a_i et b_i avec $1 \leq i \leq \tau$ sont des coefficients réels et τ correspond au décalage optimal. Si le modèle complet est significativement plus précis que le modèle restreint, il est possible de conclure que X Granger-cause Y.

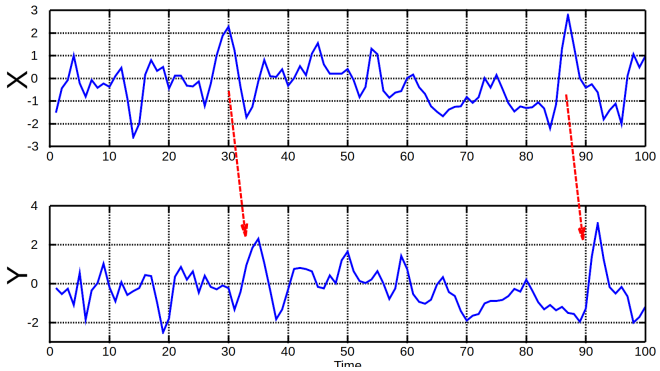


FIGURE 3.5 – **La causalité de Granger.** By BiObserver - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=33470670

3.5 L'entropie de transfert

Dans la mesure où notre méthode MIIC utilise l'information mutuelle, nous nous sommes plus particulièrement intéressés à l'entropie de transfert qui étend la causalité de Granger aux dynamiques non linéaires [45, 46].

Le transfert d'entropie permet de mesurer de façon orientée (asymétrique par rapport au temps) la quantité d'information transmise entre deux processus aléatoires. L'entropie de transfert d'un processus X vers un autre processus Y est la quantité d'incertitude qui est réduite dans les valeurs futures de Y en connaissant les valeurs de X étant donné les valeurs passées de Y.

Le transfert d'entropie peut être écrit sous la forme :

$$T_{X\to Y} = H(Y_t \mid Y_{t-1:t-\tau}) - H(Y_t \mid Y_{t-1:t-\tau}, X_{t-1:t-\tau})$$

où $H(X)$ est l'entropie de Shannon, X_t et Y_t avec $t \in \mathbb{N}$ sont deux processus aléatoires.

Le transfert d'entropie correspond à l'information mutuelle conditionnelle en conditionnant sur l'historique de la variable influencée :

$$T_{X\to Y} = I(Y_t; X_{t-1:t-\tau} \mid Y_{t-1:t-\tau})$$

Dans la mesure où il a été montré que le transfert d'entropie est équivalent à la causalité de Granger pour les processus basés sur des vecteurs auto-régressifs [46], dans la suite de ce document, nous nous référerons uniquement au transfert d'entropie pour effectuer une comparaison avec la version temporelle de MIIC.

3.6 PCMCI+

Outre PC, la principale méthode que nous avons retenue pour comparer nos résultats est PCMCI+ [47, 48], qui comme tMIIC, est une méthode de découverte causale basée sur les contraintes pour les séries temporelles. De plus, la revue détaillée récente de Assaad et al [19] montre que PCMCI+ obtient de bons résultats par rapport aux autres approches concurrentes sur les graphes fenêtrés(Fig 3.2). Enfin, tout comme tMIIC, PCMCI+ est capable de traiter les arêtes contemporaines, ce qui nous a conduit à retenir cette méthode comme une référence de comparaison lors de l'évaluation des performances de tMIIC.

PCMCI+ est l'extension aux variables contemporaines de la méthode PCMCI [49], qui est elle même une version dérivée la méthode PC adaptée au temps et utilisant des ensembles de séparation différents.

PCMCI est basée sur le principe que, considérant le fait que les parents d'une variable sont un ensemble de conditionnement suffisant qui permet d'établir une indépendance conditionnelle, il n'est pas nécessaire de conditionner sur l'ensemble du passé comme le fait par exemple la méthode FullCI. Pour PCMCI, le conditionnement s'effectue uniquement sur un ensemble qui inclut au moins les parents d'une variable, ce qui suffit à éliminer les liens douteux avec l'avantage, étant donné que ces tests sont tous de très faible dimension par rapport à FullCI ou la causalité de Granger, d'avoir un pouvoir de détection plus élevé. Le conditionnement sur les deux parents lors de l'évaluation d'une arête permet en outre de prendre en compte l'auto-corrélation et conduit à des taux de faux positifs correctement contrôlés.

La première étape de l'algorithme PCMCI (et PCMCI+) est donc de déterminer pour chaque variable j un ensemble de parents dans le passé $\widehat{\mathcal{B}}_t^-(X_t^j)$ en utilisant l'algorithme PC stable. Pour cela PCMCI considère comme point de départ un graphe laggé semi-complet, sans arête contemporaine et où chaque nœud passé ne peut avoir d'arête qu'avec des nœuds sur le dernier pas de temps.

PCMCI utilise ensuite ces ensembles de parents pour évaluer l'indépendance conditionnelle des arêtes en appliquant des tests d'indépendance conditionnelle momentanée (MCI).

$$MCI : X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i) \tag{3.1}$$

Dans PCMCI+, les sets de conditionnements sont étendus pour inclure dans $\mathcal{S}$ les nœuds contemporains :

$$X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i) \tag{3.2}$$

Enfin, le squelette obtenu est orienté selon la temporalité pour les arêtes décalées dans le temps et, dans PCMCI+, les arcs contemporains sont orientés en suivant le même principe que PC (détection de structures en V si le nœud central des triplets n'est pas dans l'ensemble de séparation des nœuds externes du triplet) et les mêmes options que dans PC sont disponibles dans PCMCI+ avec, par exemple, la règle de majorité.

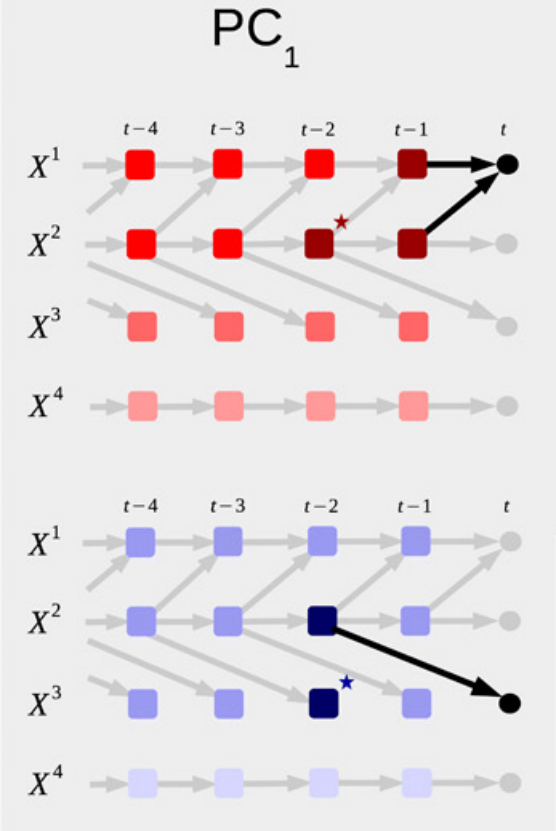


FIGURE 3.6 – PC_1, **première étape de l'algorithme PCMCI+.** Dans la première itération (p = 0), les variables indépendantes inconditionnellement (par exemple non corrélées) sont supprimées (nuances les plus claires de rouge et de bleu). Dans la deuxième itération (p = 1), les variables qui deviennent indépendantes en fonction du nœud ayant la plus grande dépendance dans l'itération précédente sont supprimées. Dans la troisième itération (p = 2), les variables indépendantes sont supprimées conditionnellement aux deux nœuds les plus forts et ainsi de suite jusqu'à ce qu'il n'y ait plus de condition à tester. De cette façon, PC_1 converge de manière adaptative vers seulement quelques conditions pertinentes (rouge et bleu foncés) qui incluent les parents causaux avec une forte probabilité et potentiellement quelques faux positifs (marqués d'une étoile).

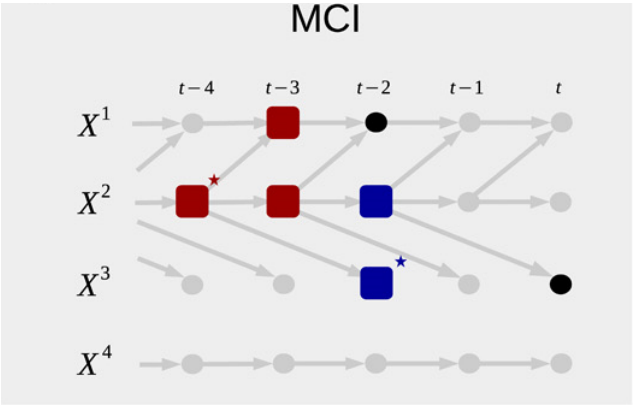


FIGURE 3.7 – **Evaluation des arêtes avec MCI.** Ces conditionnements de faibles dimensions sont utilisés dans le test d'indépendance conditionnelle MCI : pour tester $X^1_{t-2} \to X^3_t$, les conditions (cases bleues) $\widehat{\mathcal{B}}^-_t(X^3_t)$ suffisent à établir une indépendance conditionnelle, tandis que les conditions supplémentaires $\widehat{\mathcal{B}}^-_{t-2}(X^1_{t-2})$ imposées aux parents (cases rouges) tiennent compte de l'auto-corrélation et font de MCI un estimateur de la force causale. (illustration provenant de la méthode PCMCI simple)

4 Problématiques d'une version temporelle

4.1 Choix du type de graphe

La première problématique à laquelle j'ai été confronté en démarrant ce projet de version temporelle de MIIC à été le choix du type de graphe à inférer.

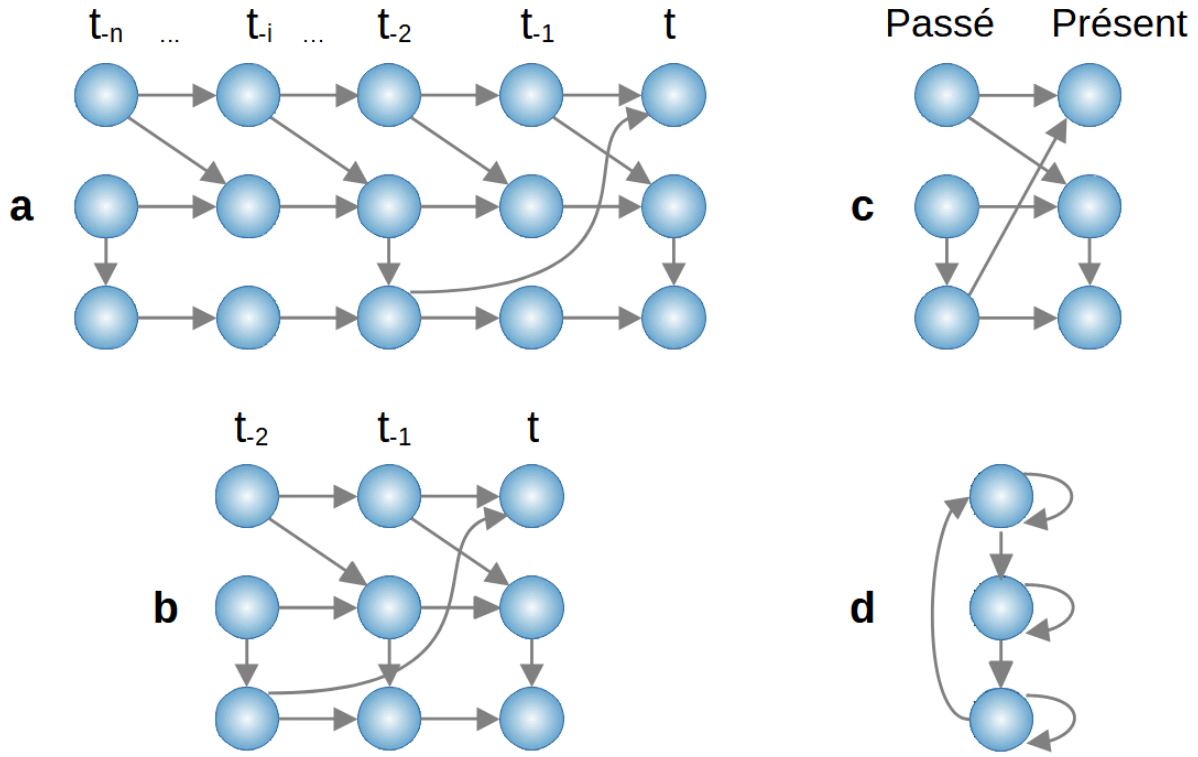


Figure 4.1 – **Types de graphes causaux** : a graphe entier, b graphe fenêtré, c graphe sommaire étendu, d graphe sommaire

- Le graphe entier, éventuellement infini, permet de rendre compte de l'ensemble de la dynamique du système à étudier mais est difficile à manipuler en pratique, le nombre de nœuds augmentant en multiple du nombre de pas de temps présents dans la série chronologique.
- Le graphe fenêtré, qui est une version réduite du graphe entier, contient uniquement le nombre de pas de temps nécessaire pour rendre compte de la dynamique du système étudié. Si ce type de représentation apparait, de prime abord, plus facile à manipuler qu'un graphe complet, il est nécessaire de déterminer a priori la taille de la fenêtre et de faire l'hypothèse de stationnarité.
- Le graphe sommaire étendu [50], quant à lui, n'essaie pas de rendre compte de la totalité de la dynamique du système puisque seuls deux pas de temps sont retenus : le passé et le présent. Ce type de graphe présente l'intérêt majeur d'un nombre de nœuds limités à 2 $*$ nombre initial de variables, ce qui limite l'es-

pace de recherche et donc le temps d'exécution mais au prix d'une information moindre.

- Le graphe sommaire [51], enfin, ne distingue pas le passé du présent, ce qui diminue encore davantage l'espace de recherche, toujours au prix d'une moindre information car la seule information restituée est de savoir si un nœud en influence un autre ou lui-même sans précision sur l'aspect temporel. Outre le peu d'information contenue dans ce type de graphe, les graphes produits peuvent comporter des cycles, ce qui les rend difficiles à inférer pour des méthodes qui supposent un graphe acyclique.

Notre objectif étant d'obtenir des graphes les plus informatifs possibles, nous avons donc retenu deux types de graphes : le graphe entier pour le cas non stationnaire et le graphe fenêtré lorsque nous posons l'hypothèse de stationnarité. A ces graphes, nous avons ajouté une représentation synthétique du graphe fenêtré pour avoir une visualisation plus simple des résultats.

4.2 Les contributeurs en fonction du temps

La reconstruction d'un graphe causal déployé dans le temps, qu'il soit entier ou fenêtré, pose la problématique d'exclure les contributeurs futurs lors de l'étape du squelette. En effet, avec l'hypothèse que le futur ne peut avoir d'effet sur le présent ou le passé, conditionner sur le futur aboutirait à des indépendances conditionnelles erronées.

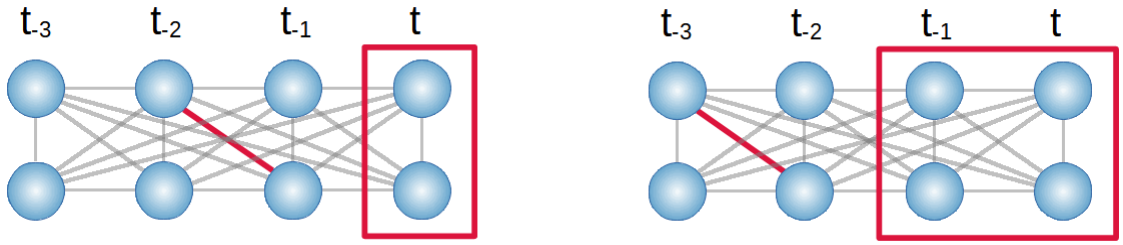


FIGURE 4.2 – **Contributeurs à exclure en fonction du temps.** Selon l'arête évaluée, les contributeurs à exclure sont différents : avec une arête $t_{-2} - t_{-1}$, les nœuds à t ne peuvent pas être des contributeurs alors que pour une arête $t_{-3} - t_{-2}$, l'ensemble des nœuds à t et t_{-1} doit être exclu.

4.3 L'orientation en fonction du temps

Le temps apportant l'hypothèse forte que le futur ne peut être la cause du présent ou du passé, une connaissance supplémentaire est donc ajoutée lors de la phase d'orientation. Cet effet du temps doit cependant être relativisé si l'on accepte la présence de variables latentes.

En effet, si on exclut par hypothèse les variables latentes, l'orientation des arêtes avec des pas de temps différents peut être totalement déterminée par la temporalité. Par contre, autoriser les variables latentes introduit une incertitude sur la queue de l'arête puisqu'il devient possible qu'une autre variable intervienne.

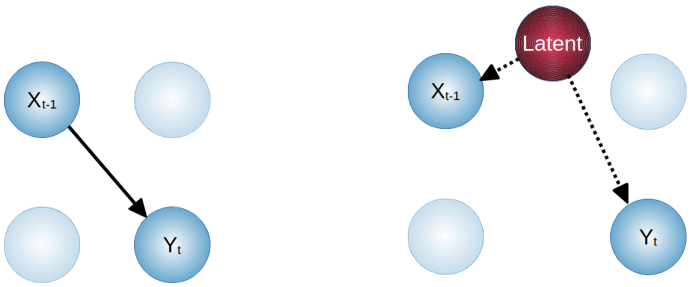


FIGURE 4.3 – **Orientation en fonction du temps.** Sans variable latente, l'orientation est déduite du temps. En cas de variable latente, l'orientation part de la variable latente vers les autres variables.

4.4 Consistance des arêtes

Si l'on fait l'hypothèse de la stationnarité vient une autre problématique sur les arêtes identiques par translation dans le temps. En effet, en utilisant un graphe entier ou fenêtré, des arêtes identiques par stationnarité pourraient être évaluées différemment par la méthode de découverte causale, produisant un graphe en violation de l'hypothèse de stationnarité.

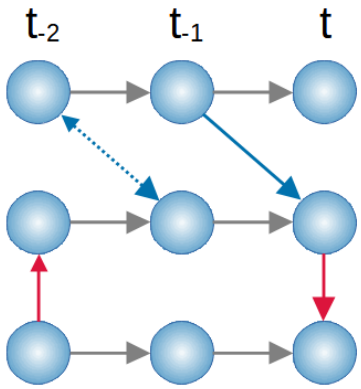


FIGURE 4.4 – **Arêtes différentes selon le temps.** Avec l'hypothèse de stationnarité, les arêtes équivalentes par translation doivent être identiques. Sur cet exemple, un post-traitement est donc nécessaire pour que le graphe résultant soit conforme à l'hypothèse de stationnarité.

5 La version temporelle stationnaire

Nous avons choisi de commencer par implémenter une version stationnaire de tMIIC. En effet, l'application cible (des séquences d'images d'environnements tumoraux reconstitués sur micro-puces, cf section 5.4) ne comprend que 36 expériences mais ces expériences ayant entre 720 et 1440 pas de temps, le nombre de pas de temps utilisables est de 46 935 en posant l'hypothèse de stationnarité.

5.1 tMIIC stationnaire

L'objectif en mode stationnaire étant d'établir un graphe fenêtré déplié dans le temps, nous avons introduit un paramètre d'entrée τ qui détermine le nombre de pas de temps dans le passé sur lequel s'effectue la découverte causale. Ce paramètre a ensuite été utilisé pour pré-processer les variables et les données.

5.1.1 Pré-processing

Décalage des variables

Pour établir le graphe déplié dans le temps, une première étape a donc été de dupliquer les variables en les décalant sur $\tau + 1$ pas de temps.

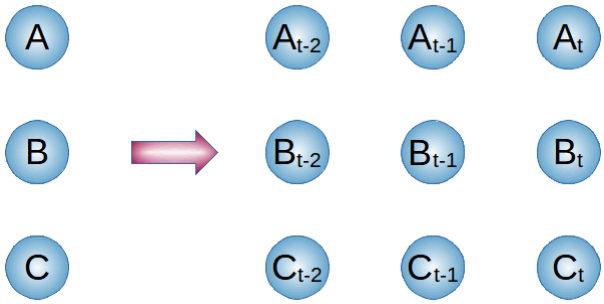


FIGURE 5.1 – **Décalage des variables.** De nouvelles variables sont générées sur la fenêtre $t - \tau$ à t.

Décalage des données

De la même manière, les données d'entrée, qui sont les valeurs prises par les variables aux pas de temps successifs ont été dupliquées et décalées dans le temps. En notant l la longueur des trajectoires, chaque trajectoire initiale permet donc de créer $l - \tau$ échantillons.

A ce stade, une remarque s'impose liée à la corrélation des échantillons inhérente aux séries chronologiques : dans le monde réel, une corrélation entre chaque variable et son passé est souhaitable car il est attendu que chaque objet soit dépendant de son propre passé. Cependant, dans une découverte causale non temporelle, il est nécessaire de poser l'hypothèse d'indépendance entre échantillons pour que l'information contenue dans les données corresponde bien au nombre total d'échantillons. En cas d'échantillons corrélés, une correction est nécessaire avec un nombre effectif d'échantillons (paramètre Neff de la méthode MIIC [36]). Dans une série temporelle, fortement

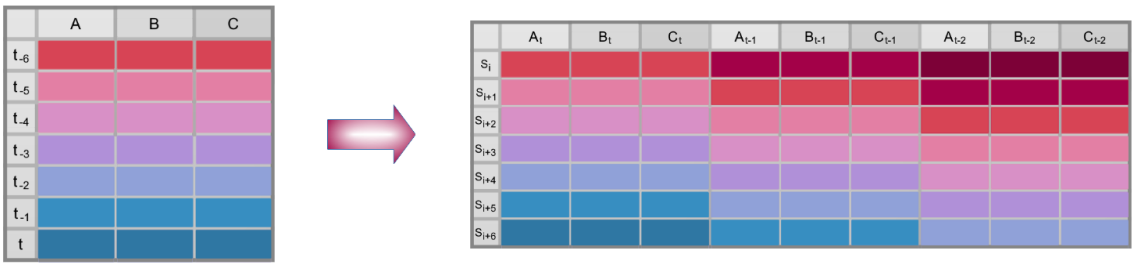


FIGURE 5.2 – **Décalage des données**, à partir de l pas de temps, nous obtenons $l - \tau$ échantillons décalés dans le temps.

corrélée, nous devrions donc avoir une information moindre que le nombre de pas de temps total. Cependant, dans un graphe déplié dans le temps, cette correction du nombre d'échantillons n'est pas nécessaire [49, 52-54] car le conditionnement sur les valeurs précédentes de chaque variable impliqué par un graphe déplié dans le temps assure l'indépendance de chaque échantillon conditionné sur les valeurs précédentes, comme donné par la vraisemblance,

$$\begin{aligned} q(\boldsymbol{X_1}, \boldsymbol{X_2}, ..., \boldsymbol{X_N}) &= q(\boldsymbol{X_N}|\boldsymbol{X_{N-1}}, ..., \boldsymbol{X_1})\, q(\boldsymbol{X_{N-1}}|\boldsymbol{X_{N-2}}, ..., \boldsymbol{X_1}) \cdots q(\boldsymbol{X_1}) \\ &\simeq q(\boldsymbol{X_N}|\boldsymbol{X_{N-1}}, ..., \boldsymbol{X_{N-\tau}})\, q(\boldsymbol{X_{N-1}}|\boldsymbol{X_{N-2}}, ..., \boldsymbol{X_{N-1-\tau}}) ... \end{aligned}$$

5.1.2 Elagage du squelette initial

Comme nous l'avons vu dans la section 4.2, pour la phase squelette en mode stationnaire, nous avons besoin de prendre en compte les problématiques de listes de contributeurs variables en fonction du temps et de consistance des arêtes.

Ces deux problématiques ont trouvé une réponse par une réduction du squelette initial : plutôt que de commencer d'un graphe complet et ensuite devoir gérer des listes de contributeurs variables et vérifier que les arêtes sont consistantes, l'ensemble des arêtes identiques par stationnarité et n'ayant pas de nœud sur le denier pas de temps a été retiré du graphe.

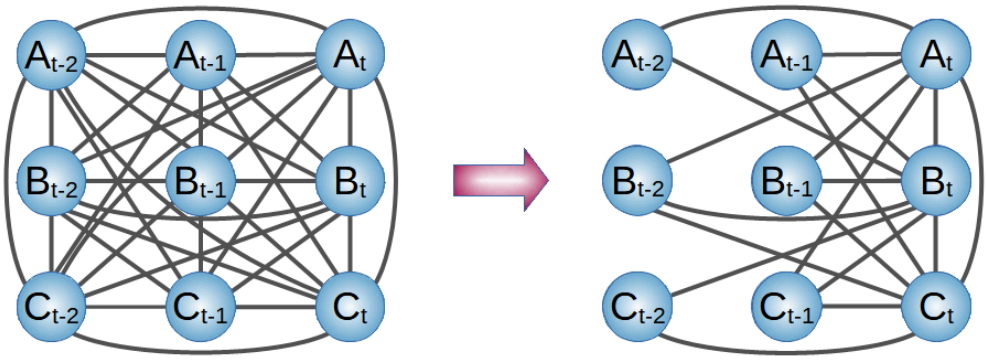


FIGURE 5.3 – **Elagage des arêtes dupliquées.** En assumant la stationnarité, le graphe initial peut être élagué en supprimant les arêtes identiques décalées dans le temps.

Cette élagage préalable permet de ne conserver qu'une version des arêtes identiques par stationnarité, ce qui élimine de facto la gestion de la consistance des arêtes

dupliquées. Il permet également d'éviter de gérer des listes de contributeurs variables car les arêtes ayant toujours un nœud sur le dernier pas de temps, il n'est pas nécessaire de limiter les contributeurs selon les pas de temps futurs. Avantage supplémentaire, l'élagage initial réduit le nombre d'arêtes à évaluer et donc le temps d'exécution.

Les règles d'inférence classique de MIIC sont ensuite appliquées pour déduire le squelette : suppression des arêtes en cas d'indépendance non conditionnelle, choix des meilleurs contributeurs, suppression des arêtes en cas d'indépendance conditionnelle avec les meilleurs contributeurs possibles.

Pour préparer la phase d'orientation, les arêtes restantes sont dupliquées en utilisant l'hypothèse de stationnarité pour permettre une identification correcte des triplets ouverts.

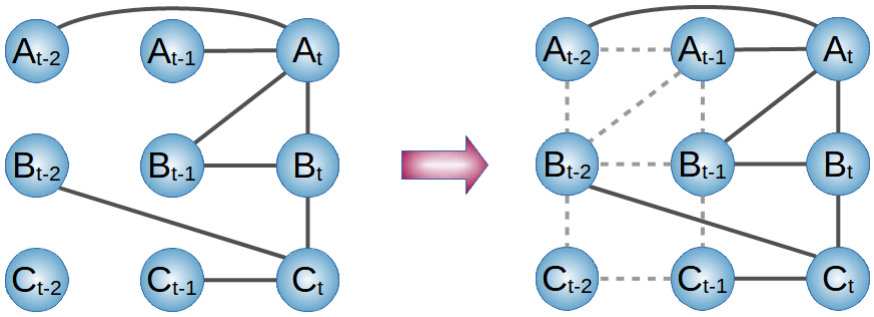


FIGURE 5.4 – **Duplication des arêtes par stationnarité.** Certains triplets apparaissant ouverts sur le graphe élagué, par exemple $A_{t-1} - A_t - B_{t-1}$, ne peuvent être utilisés pour la phase d'orientation car, du fait de la stationnarité, ils ne sont pas ouverts. Le graphe correct pour la phase d'orientation est obtenu après duplication par stationnarité, et, pour cet exemple, l'arête $A_{t-1} - B_{t-1}$ vient fermer le triplet $A_{t-1} - A_t - B_{t-1}$.

5.1.3 Orientation selon le temps

Si tMIIC inclut bien sûr le temps lors de la phase d'orientation, cependant, contrairement à MIIC, seule une partie des triplets ouverts va être retenue pour la phase d'orientation. En effet, il est inutile d'essayer d'orienter les triplets passés (n'ayant aucun nœud au dernier pas de temps) car ces triplets sont des versions dupliquées de triplets postérieurs.

Pour inclure le temps dans l'orientation tout en conservant les capacités de MIIC à déduire l'orientation depuis la signature de causalité pour les arêtes contemporaines ainsi que la découverte de variables latentes, les règles d'orientation classiques de MIIC ont été conservées mais avec une initialisation modifiée (figure 5.6).

En effet, dans le mode classique sans le temps, les extrémités des arêtes sont initialisées avec une probabilité de 0.5 d'être une tête (une conséquence) et les arêtes n'appartenant pas à un triplet ouvert ne peuvent être orientées. Dans la version temporelle, les orientations des arêtes avec décalage dans le temps sont initialisées avec une orientation pointant vers le futur, indiquant que nous sommes certains que cette extrémité, en accord avec la temporalité, ne peut pas être une cause de la variable dans le passé. L'initialisation de la partie opposée de l'arête (le côté du pas de temps le plus ancien) varie selon que nous autorisons ou pas les variables latentes. Sans variable latente, la probabilité d'orientation est 0, indiquant que cette variable n'est pas causée

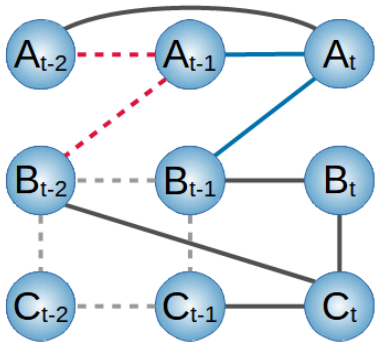


FIGURE 5.5 – **Triplets inclus/exclus de la phase d'orientation.** Sur cet exemple, le triplet rouge $A_{t-2} - A_{t-1} - B_{t-2}$ n'a pas besoin d'être orienté car il peut être déduit par stationnarité du triplet bleu $A_{t-1} - A_t - B_{t-1}$.

(par hypothèse) par une variable latente non-observée dans le jeu de données. Lorsque les variables latentes sont autorisées, la valeur initiale de la probabilité d'orientation de cette extrémité est fixée à 0.5, ce que traduit notre incertitude initiale sur la présence d'une cause commune latente.

tMIIC présente une autre particularité : les arêtes n'appartenant pas à des triplets ouverts, qui sont, dans la version classique, laissées non orientées reçoivent dans la version temporelle une orientation vers le futur.

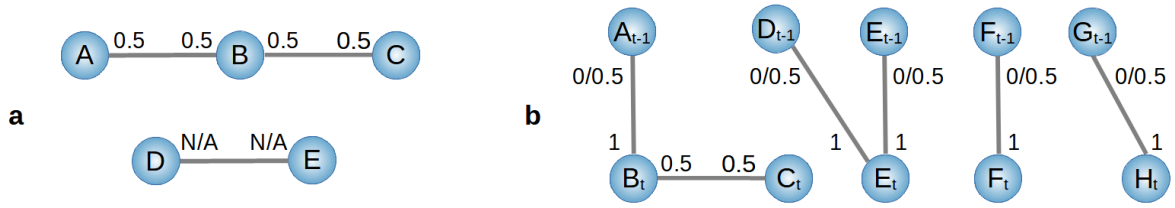


FIGURE 5.6 – **Probabilités initiales affectées aux arêtes.** a Dans le mode classique sans le temps, les extrémités sont initialisées avec une probabilité de 0.5. b En temporel, les arêtes avec un décalage dans le temps ont une initialisation de leur orientation vers le futur, y compris sans triplet ouvert.

L'algorithme classique d'orientation des arêtes est ensuite appliqué pour inférer les orientations selon la causalité, ce qui permet de déceler les éventuelles variables latentes, distinguer les arêtes causales putatives ou authentiques ainsi que l'orientation des arêtes contemporaines pour parvenir au graphe final.

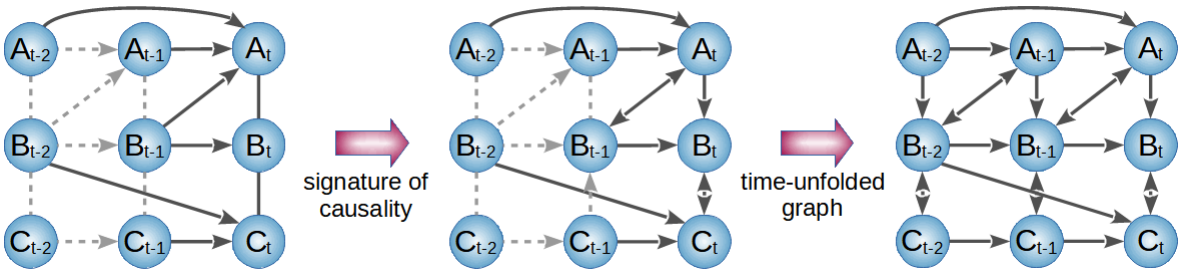


FIGURE 5.7 – **Phase d'orientation** : le squelette est pré-orienté en fonction du temps, puis l'algorithme utilise la signature de causalité pour établir le graphe final déployé dans le temps

5.1.4 Conservation des fonctionnalités de la version non temporelle

tMIIC prend en charge l'ensemble des fonctionnalités disponibles dans MIIC, qui ont toutes été adaptées si nécessaire dans la version temporelle. Il y a par exemple la gestion des variables latentes, décrite ci-dessus mais il est possible de citer également la fonctionnalité de variable contextuelle.

Dans le cadre temporel, ces variables contextuelles ont été traduites par une variable constante tout au long de chaque trajectoire de la série temporelle. Contrairement aux autres variables, les variables contextuelles ne sont pas dépliées dans le temps puisque ces variables sont constantes pour chaque expérience. Autre spécificité des variables contextuelles, lorsque nous évaluons l'indépendance conditionnelle d'une arête où un des nœuds est une variable contextuelle, l'historique de l'autre nœud (variable non contextuelle) est exclu de la liste des contributeurs possibles.

5.1.5 Fonctionnalités supplémentaires

Deux fonctionnalités spécifiques à la version temporelle ont été développées :

Moyenne mobile

Cette option peut être utile dans les cas où les données comportent des fluctuations transitoires. Pour illustrer l'intérêt de cette fonctionnalité, on peut prendre un exemple spécifique de ce type de données : les tests journaliers de dépistage du Covid-19. Vu que le nombre de tests journaliers de dépistage du Covid-19 est moins important le weekend, un lissage des données sur une semaine permet d'avoir une meilleure estimation de l'évolution du nombre de contaminations.

Intervalle entre deux pas de temps

Dans le cas où le nombre maximum de pas de temps à considérer pour rendre compte de la dynamique devient trop important (nombre total de nœuds dans le réseau = (nombre de pas temps passés + 1) * nombre de variables en entrée) pour être calculable, il peut être intéressant d'augmenter le délai entre les différents pas de temps. Pour cela, un paramètre δ_t a été introduit. Par exemple, un réseau de 10 variables avec 50 pas de temps d'historique génèrerait un graphe déplié dans le temps de 550 noeuds, ce qui peut devenir difficile à calculer. En ajoutant un δ_t de 10, la taille du graphe final diminue à 60 nœuds et devient nettement plus accessible en terme de temps de calcul.

Si le δ_t est utile pour réduire la taille du graphe et le temps de calcul, il présente le désavantage de diminuer l'information utile puisque le décalage des données ne va également conserver qu'une ligne tous les δ_t. tMIIC ajuste automatiquement le nombre d'échantillons effectifs lors de l'utilisation du δ_t et, par exemple, avec un δ_t de 10, le paramètre N_{eff} sera ajusté au nombre total de pas de temps / 10.

L'introduction du paramètre δ_t nous a conduit à reconsidérer les paramètres à fournir à tMIIC et nous avons remplacé le paramètre τ initial par le nombre de couches λ et δ_t, le nombre de pas entre deux couches.

5.1.6 Algorithme global

Algorithm 4
Reconstruction causale par tMIIC stationnaire (sans variable latente)

Require: l the number of layers and δ_t the number of time steps between layers
- Preprocessing
Duplicate variables and data along $l-1$ layers back in history, separating each layer by δ_t time steps
- Skeleton reconstruction
$\mathcal{G} \leftarrow$ the complete graph on V
for all edges $X_t - Y_{t'} \in \mathcal{G}$ **do**
  if ((t is past and t' is past) or I'$(X_t; Y_{t'})) \leqslant 0$ **then**
    Delete edge $X_t - Y_{t'}$ from $\mathcal{G}$
    Sepset$\{X_t, Y_{t'}\} \leftarrow \emptyset$
  else
    Find most contributing node $Z_{t''} \in \{\text{adj}(X_t) \cup \text{adj}(Y_{t'})\}$ which maximizes $R(X_t, Y_{t'}; Z_{t''}|\emptyset)$
  end if
end for
while There is a link $X - Y$ with $R(X, Y; Z|\{U_i\}) > 1/2$ **do**
  for Top link $X - Y$ with highest rank $R(X, Y; Z|\{U_i\})$ **do**
    Expand contributing set $\{U_i\} \leftarrow \{U_i\} + Z$
    if I'$(X; Y|\{U_i\}) \leqslant 0$ **then**
      Delete edge $X - Y$ from $\mathcal{G}$
      Sepset$\{X, Y\} \leftarrow \{U_i\}$
    else
      Find next most contributing node $Z \in \{\text{adj}(X) \cup \text{adj}(Y)\}$ and compute $R(X, Y; Z|\{U_i\})$
    end if
    Sort the rank list $R(X, Y; Z|\{U_i\})$
  end for
end while
Duplicate edges assuming stationnarity
- Skeleton orientation
Initialize $X_t - Y_{t'}$ as $X_t \rightarrow Y_{t'}$ if $t < t'$ and $X_t \leftarrow Y_{t'}$ if $t > t'$
Sort list of unshielded triples $\mathcal{L}_c = \{(X_t, Z_{t'}, Y_{t''})_{X_t \neq Y_{t''}} \mid t \text{ or } t' \text{ or } t'' \text{ is present}\}$ in decreasing order of $|\text{I}'(X_t, Z_{t'}, Y_{t''}|\{U_i\})|$
repeat
  Take $(X, Z, Y)_{X \neq Y} \in \mathcal{L}_c$ with highest $|\text{I}'(X; Y; Z|\{U_i\})|$ on which orientation rules can be applied
  if I'$(X; Y; Z|\{U_i\}) < 0$ **then**
    if $(X, Z, Y)_{X \neq Y}$ has no diverging orientation, orient as $X \rightarrow Z \leftarrow Y$
  else if I'$(X; Y; Z|\{U_i\}) > 0$ **then**
    if $(X, Z, Y)_{X \neq Y}$ has one converging orientation, propagate orientation as $X \rightarrow Z \rightarrow Y$
  end if
  Update all orientations of $(X, Z, Y)_{X \neq Y} \in \mathcal{L}_c$
until No additional orientation can be obtained
return $\mathcal{G}$

5.1.7 Mise à disposition de la communauté scientifique

Via le dépôt public MIIC sur Github

L'ensemble de l'algorithme décrit ci-dessus a été implémenté dans le package R de MIIC et est accessible à tous depuis un dépôt Github à l'adresse `https://github.com/miicTeam/miic_R_package/tree/tmiic15x`.

Via un serveur en ligne

De plus, notre laboratoire ayant pour objectif de rendre l'ensemble des outils que nous développons aisément utilisable par la communauté scientifique, tMIIC est disponible par l'intermédiaire d'un serveur en ligne `https://miic.curie.fr/workbench_timeseries.php`.

Outre des modifications du cœur du package R, l'implémentation de tMIIC stationnaire a donc été réalisée aussi sur le serveur qui héberge la version classique de MIIC, en adaptant l'ensemble des programmes, que ce soit côté front-end ou back-end. Côté front-end, les pages web ont ainsi été complétées pour présenter les nouvelles fonctionnalités de tMIIC.

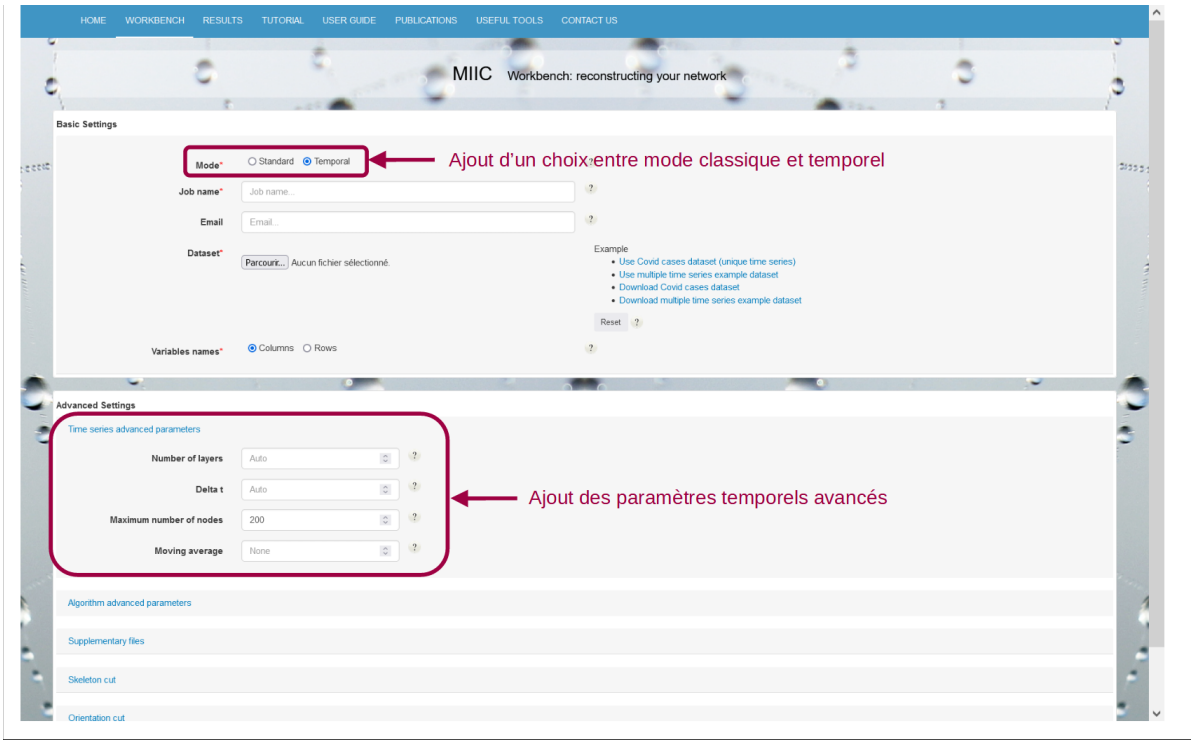


FIGURE 5.8 – **Page de lancement de travaux** : le nouveau mode a été ajouté avec ses paramètres dédiés

L'ensemble du back-end a bien sûr été modifié en conséquence pour prendre en compte ces nouveaux paramètres, les transférer au package R pour inférer le réseau causal attendu et préparer l'affichage des résultats.

Les modifications les plus importantes et les plus longues côté serveur ont sans aucun doute été celles relatives à la visualisation des réseaux obtenus. En effet, initialement, le module d'affichage des graphes causaux ne prévoyait pas de dessiner des

boucles ou de multiples arêtes entre deux nœuds, puisque ce type de cas n'était pas possible sur des graphes non temporels. De plus la représentation des graphes fenêtrés à nécessité la mise en œuvre de layouts spécifiques permettant d'aligner les nœuds des variables à travers le temps.

Au final, de multiples visualisations ont été ajoutées au serveur spécifiques aux réseaux temporels :

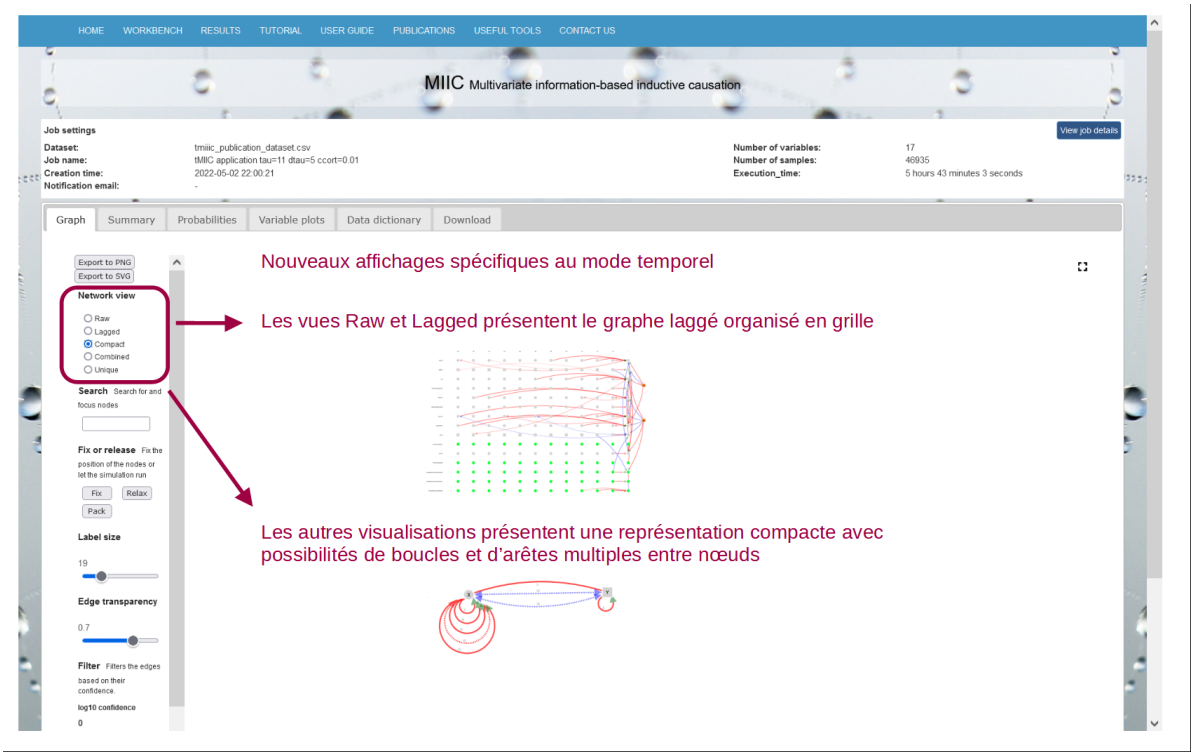


FIGURE 5.9 – **Page de visualisation des réseaux causaux obtenus par tMIIC** : des représentations en grille avec l'ensemble des nœuds décalés dans le temps et compactes avec des boucles et de multiples arêtes entre nœuds ont été développées pour visualiser aisément les résultats.

5.2 Relation à la causalité de Granger-Schreiber

Comme exposé dans le chapitre 3, le concept de causalité temporelle a été initialement formulé par Granger sans référence à aucun modèle graphique en comparant l'auto-régression avec ou sans valeurs passées des variables causales possibles. Ce concept a ensuite été étendu aux relations non linéaires par Schreiber en utilisant la notion d'entropie de transfert, $T_{X\to Y}$, qui peut être exprimée en termes d'informations conditionnelles multivariées,

$$T_{X\to Y} = I(Y_t; \boldsymbol{X}_{t'<t} | \boldsymbol{Y}_{t'<t}) \tag{5.1}$$

où $\boldsymbol{X}_{t'<t}$ et $\boldsymbol{Y}_{t'<t}$ désignent les ensembles de variables, $X_{t'}$ et $Y_{t'}$, prises à des instants t' antérieurs à t.

Alors que l'Eq. 5.1 est asymétrique sur la permutation X/Y, une simple comparaison de l'asymétrie d'entropie de transfert (*par exemple* $T_{X\to Y} > T_{Y\to X} \geqslant 0$) ne se traduit pas nécessairement par une direction causale car cette asymétrie est également attendue pour des relations non causales. Fait intéressant, c'est en fait l'absence d'entropie de transfert dans une direction (*par exemple* $T_{Z\to X} \approx 0$) qui suggère la possibilité d'une relation causale dans une direction opposée, $X \to Z$, comme dans le cas des structures en V dans les méthodes de découverte causale basées sur les graphes, à condition qu'une cause commune latente puisse être exclue entre les deux variables.

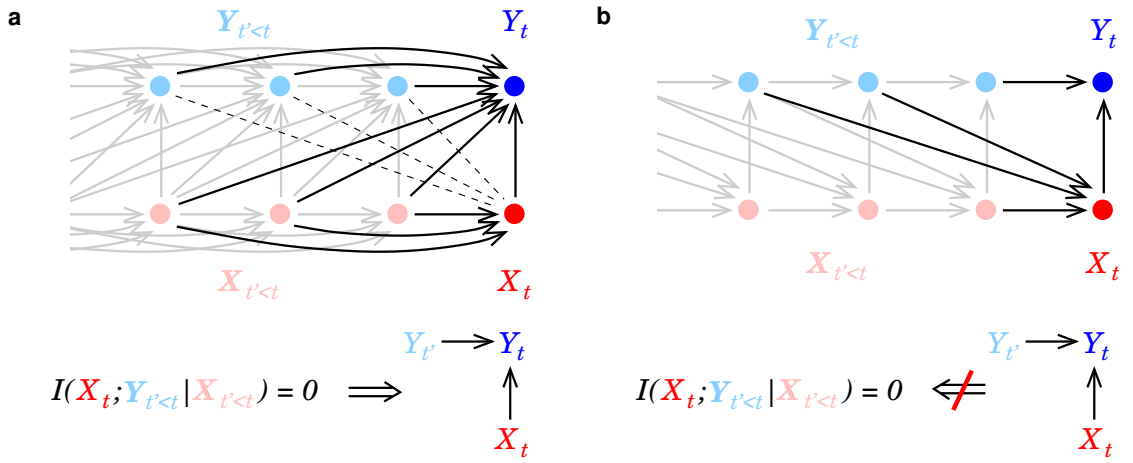


FIGURE 5.10 – **Réseau causal déplié dans le temps et relation avec la causalité temporelle de Granger-Schreiber. a**, une disparition de l'entropie de transfert, c'est-à-dire $T_{Y\to X} = I(X_t; \boldsymbol{Y}_{t'<t} | \boldsymbol{X}_{t'<t}) = 0$, implique *i)* l'absence d'arête (en tirets) entre X_t et tout $Y_{t'}$, avec $t' < t$, et *ii)* si X_t est adjacent à Y_t, la présence de structures en V temporelles (2 variables + temps), $Y_{t'} \to Y_t \leftarrow X_t$, pour tout $Y_{t'}$ adjacent à Y_t, avec $t' < t$ (Théorème 1). Ces résultats peuvent être facilement étendus pour inclure la présence d'autres variables observées, $\boldsymbol{V}_{t'\leqslant t}$, en redéfinissant l'entropie de transfert comme, $T_{Y\to X} = I(X_t; \boldsymbol{Y}_{t'<t} | \boldsymbol{X}_{t'<t}, \boldsymbol{V}_{t'\leqslant t})$, ce qui écarte les contributions de chemins indirects à travers d'autres variables observées, $\boldsymbol{V}_{t'\leqslant t}$. **b**, En revanche, la présence d'une structure en V temporelle (2 variables + temps), $Y_{t'} \to Y_t \leftarrow X_t$ n'implique pas une entropie de transfert nulle, tant qu'il reste une arête entre un $Y_{t''<t}$ et X_t.

Nous clarifions avec le théorème 1 ci-dessous cette relation entre la causalité temporelle sans référence à aucun modèle structurel (Eq. 5.1) et la causalité structurelle associée aux modèles graphiques causaux dépliés dans le temps ($\mathcal{G}_t$). Cela met en lumière les fondements communs des causalités temporelle et structurelle au-delà de leurs définitions apparemment sans rapport.

Théorème 1. [$T_{Y\to X} = 0$ implique des structures en V temporelles ($2\ var + t$)]
Si X_t est adjacent à Y_t dans $\mathcal{G}_t$ et $T_{Y\to X} = I(X_t; \boldsymbol{Y}_{t'<t} | \boldsymbol{X}_{t'<t}) = 0$, alors pour tout $Y_{t'}$ adjacent à Y_t dans $\mathcal{G}_t$, avec $t' < t$, il existe une structure en V temporelle (2 variables + temps), $Y_{t'} \to Y_t \leftarrow X_t$, dans $\mathcal{G}_t$.

Preuve : Nous raisonnons par contradiction en nous basant sur la Fig.5.10 ci-dessus.
S'il existe $Y_{t'}$ adjacent à Y_t tel que $Y_{t'} - Y_t - X_t$ n'est pas une structure en V, alors $T_{Y\to X} = I(X_t; \boldsymbol{Y}_{t'<t} | \boldsymbol{X}_{t'<t}) \neq 0$, comme $Y_t \notin \boldsymbol{X}_{t'<t}$ ou X_t est adjacent à $Y_{t'}$.

Il faut noter cependant que la réciproque du théorème 1 n'est pas vraie : une structure en V temporelle n'implique pas une entropie de transfert nulle, comme indiqué avec le contre-exemple dans la Fig. 5.10b. En conséquence, la présence d'une structure en V temporelle, $Y_{t'} \to Y_t \leftarrow X_t$ dans $\mathcal{G}_t$, n'implique pas nécessairement une entropie de transfert nulle, $T_{Y\to X} = 0$, tant qu'il reste une arête entre un $Y_{t''}$ et X_t, comme dans l'exemple de la figure.

Ainsi, la causalité de Granger-Schreiber est en fait trop restrictive et peut passer à côté d'effets causaux réels, qui peuvent être découverts par des méthodes de découverte causale structurelle comme tMIIC. De plus, la causalité de Granger-Schreiber est également connue pour inférer de fausses associations causales en excluant la présence de causes communes latentes a priori. A l'inverse, tMIIC comprend des effets décalés dans le temps ainsi que des effets synchrones provenant de variables latentes non observées.

5.3 Benchmark

5.3.1 Performance

Pour évaluer la performance de tMIIC, nous avons procédé à de multiples benchmarks et, pour comparer les résultats par rapport aux autres méthodes de découverte causale basées sur les contraintes, nous avons choisi le package Tigramite [49], qui propose de nombreuses méthodes adaptées aux séries temporelles, PC, FCI, MCI FullCI, BivCI, PCMCI et PCMCI+ ainsi que plusieurs noyaux pour tester l'indépendance conditionnelle : Partial Correlation, CMIknn et GPDC.

Deux de ces méthodes nous ont particulièrement intéressées, car capables d'orienter les arêtes contemporaines : PC et PCMCI+, ce sont donc ces deux méthodes que nous avons principalement utilisées pour comparer avec les résultats de tMIIC.

D'après l'application cible, nous avons construit un premier benchmark allant jusqu'à 100 000 pas de temps avec 15 variables et des fonctions similaires à celles présentées dans le package Tigramite, Table 5.1. Le modèle a été généré aléatoirement en utilisant les paramètres suivants :

- Nombre de variables : 15
- Nombre de pas de temps : 3 (2 pas de temps passés + présent)
- Degré moyen : 5
- Lien vers son propre passé d'une variable à $t-1$: 100 %
- Lien vers son propre passé d'une variable à $t-2$: 20 %

- Probabilité des décalages entre deux variables différentes, pas de décalage : 20 %, décalage de 1 pas de temps : 40 %, décalage de 2 pas de temps : 40 %
- Force minimale d'un lien : 0.1
- Force maximale d'un lien : 0.6
- Probabilité du bruit entre une variable et son passé : 100 % additif
- Probabilité du bruit entre variables différentes : 80 % additif, 20 % multiplicatif

TABLE 5.1 – Modèle du benchmark avec 15 nœuds.

Nœuds

$$
\begin{aligned}
X_t^1 &\leftarrow -0.47\, f_2(X_{t-1}^1) + 0.29\, f_3(X_{t-1}^2) \times \eta_1 \\
X_t^2 &\leftarrow 0.49\, f_2(X_{t-1}^2) + 0.4\, f_1(X_{t-2}^1) + \eta_2 \\
X_t^3 &\leftarrow 0.56\, f_1(X_{t-1}^3) + 0.44\, f_4(X_{t-2}^4) - 0.26\, f_2(X_{t-2}^{10}) + 0.56\, f_2(X_t^4) + \eta_3 \\
X_t^4 &\leftarrow 0.24\, f_3(X_{t-1}^4) - 0.24\, f_2(X_{t-2}^6) - 0.12\, f_4(X_{t-1}^{14}) \times \eta_4 \\
X_t^5 &\leftarrow -0.39\, f_3(X_{t-1}^5) - 0.42\, f_3(X_{t-2}^5) - 0.39\, f_3(X_t^{11}) + \eta_5 \\
X_t^6 &\leftarrow -0.32\, f_2(X_{t-1}^6) + \eta_6 \\
X_t^7 &\leftarrow -0.17\, f_4(X_{t-1}^7) - 0.17\, f_1(X_{t-2}^7) + \eta_7 \\
X_t^8 &\leftarrow 0.39\, f_4(X_{t-1}^8) - 0.46\, f_4(X_{t-1}^7) - 0.39\, f_3(X_{t-1}^1) - 0.4\, f_3(X_{t-2}^{12}) + \eta_8 \\
X_t^9 &\leftarrow -0.34\, f_1(X_{t-1}^9) + 0.43\, f_3(X_{t-2}^{12}) + \eta_9 \\
X_t^{10} &\leftarrow 0.2\, f_1(X_{t-1}^{10}) + 0.18\, f_4(X_{t-2}^9) + 0.17\, f_1(X_{t-1}^9) + 0.48\, f_3(X_{t-1}^7) - 0.26\, f_4(X_{t-1}^4) + \eta_{10} \\
X_t^{11} &\leftarrow 0.41\, f_2(X_{t-1}^{11}) + 0.54\, f_3(X_t^2) - 0.55\, f_2(X_t^{12}) + \eta_{11} \\
X_t^{12} &\leftarrow -0.45\, f_2(X_{t-1}^{12}) - 0.43\, f_4(X_{t-2}^3) - 0.17\, f_4(X_{t-2}^9) \times \eta_{12} \\
X_t^{13} &\leftarrow 0.45\, f_3(X_{t-1}^{13}) + \eta_{13} \\
X_t^{14} &\leftarrow 0.28\, f_2(X_{t-1}^{14}) + 0.37\, f_1(X_{t-2}^{12}) \times \eta_{14} \\
X_t^{15} &\leftarrow 0.52\, f_3(X_{t-1}^{15}) + \eta_{15}
\end{aligned}
$$

Fonctions

$$
\begin{aligned}
f_1(x) &= x \\
f_2(x) &= x(1 - 4e^{-\frac{x^2}{2}}) \\
f_3(x) &= x(1 - 4x^3 e^{-\frac{x^2}{2}}) \\
f_4(x) &= \cos(x)
\end{aligned}
$$

Bruits

Les η sont des bruits blancs générés pour chaque nœud ou contribution en utilisant une distribution normale :

$$\eta \sim \mathcal{N}(0,1)$$

A partir de ce modèle, dix jeux de données ont été générés avec dix initialisations du générateur aléatoire différentes. Des reconstructions de réseaux ont été effectuées sur ces dix jeux de données avec tMIIC, PC et PCMCI+. Pour PC et PCMCI+, les trois noyaux disponibles dans le package Tigramite ont été testés. Toutes les méthodes ont été lancées avec leurs paramètres par défaut à une exception : pour tMIIC, le nombre de threads à été réglé sur 8 pour obtenir une comparaison juste avec Tigramite dont les noyaux utilisent le processeur à 100 %.

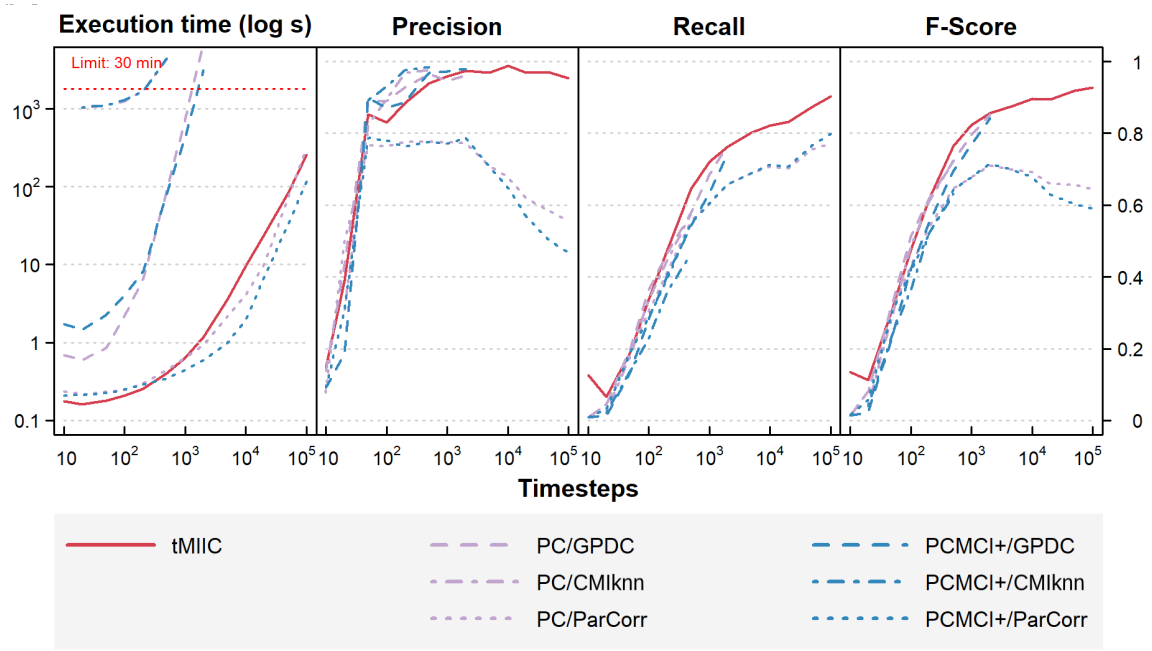


FIGURE 5.11 – **Evaluation par benchmark de la découverte causale tMIIC sur des séries temporelles générées**

Les méthodes/noyaux ont été testés sur un nombre progressif de pas de temps avec une limite à 30 minutes de temps d'exécution maximum pour maintenir la durée totale du benchmark acceptable.

Les temps d'exécution et les scores (Précision, Rappel, F-score), moyennés sur les 10 jeux de données ont été comparés entre tMIIC et les méthodes PC et PCMCI+ avec les différents noyaux (GPDC, CMIknn, ParCorr). Il apparait que tMIIC est équivalent avec les scores de PC et PCMCI+ utilisant les noyaux GPDC et CMIknn mais tMIIC est plus rapide de plusieurs ordres de grandeur. Seul le noyau ParCorr égale la vitesse d'exécution de tMIIC mais avec des scores significativement inférieurs sur des grandes tailles d'échantillons, Fig. 5.11.

En terme de reconstruction, à partir de 1 000 pas de temps, le F-score obtenu par tMIIC dépasse 0.82 et est proche de 0.9 pour 10 000 pas de temps.

En analysant plus en détail le modèle précédent, il nous est apparu que les fonctions étaient plutôt linéaires par partie et vu que, dans le cadre d'application à des données du monde réel, nous ne pouvons pas assumer que les relations sont linéaires,

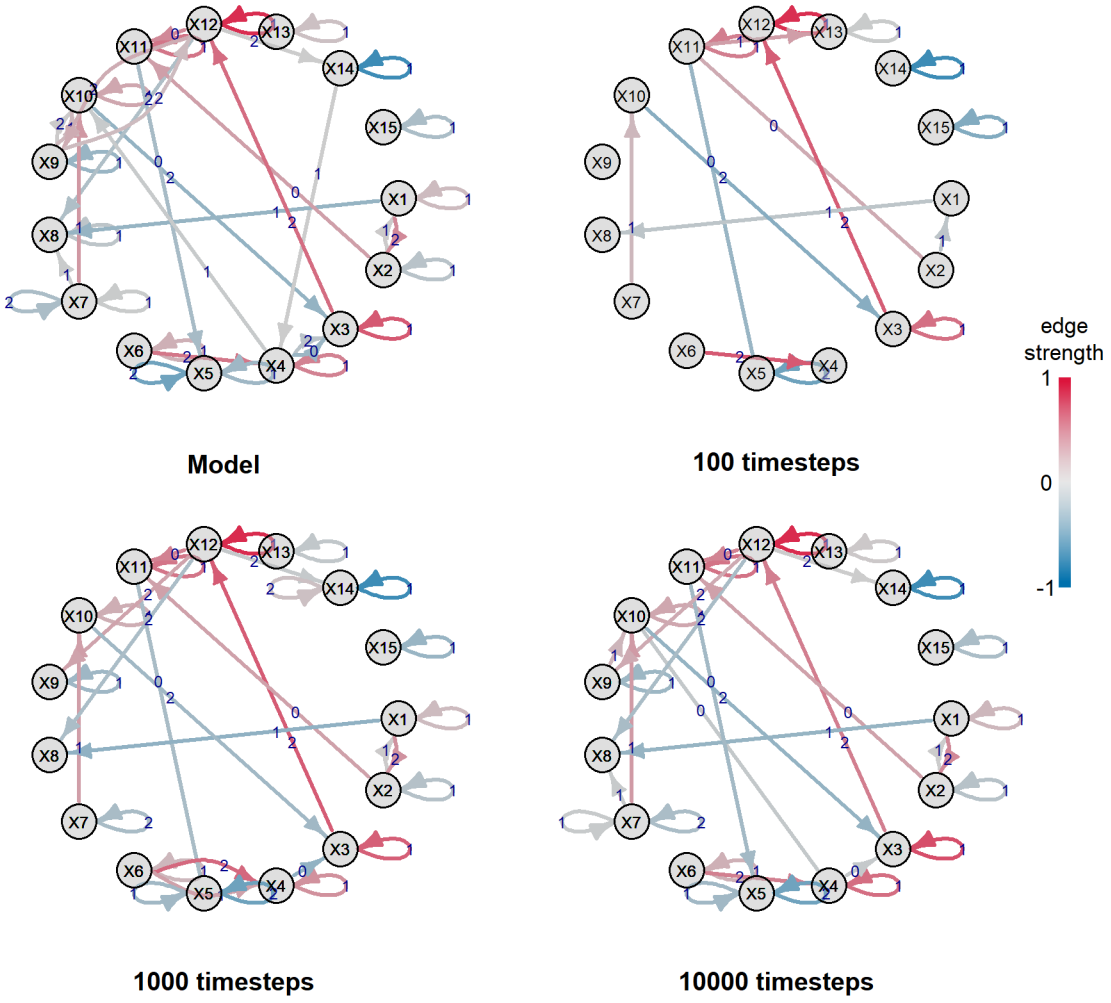


FIGURE 5.12 – **Reconstructions obtenues par tMIIC sur des séries temporelles générées** avec le modèle initial et les réseaux obtenus pour 100, 1 000 et 10 000 pas de temps. "Edge strength" est le coefficient de corrélation de Spearman, éventuellement partiel s'il y a des contributeurs.

nous avons construit un deuxième modèle en introduisant la possibilité de combinaisons non-linéaires entre variables, Table 5.2. Ce modèle a lui aussi été généré aléatoirement avec les paramètres suivants :

- Nombre de variables : 15
- Nombre de pas de temps : 3 (2 pas de temps passés + présent)
- Degré moyen : 5
- Lien vers son propre passé d'une variable à $t-1$: 90 %
- Lien vers son propre passé d'une variable à $t-2$: 20 %
- Probabilité des décalages entre deux variables différentes, aucun décalage : 20 %, décalage de 1 pas de temps : 50 %, décalage de 2 pas de temps : 30 %
- Force minimale d'un lien : 0.5
- Force maximale d'un lien : 0.9
- Type de bruit des contributions : un lien multiplicatif, les autres additifs
- Nombre de combinaisons : 4

TABLE 5.2 – Modèle à 15 nœuds avec combinaisons des contributeurs.

Noeuds

$$
\begin{aligned}
X_t^1 &\leftarrow \eta - 0.7\, f_6(u(\eta + X_{t-1}^1)) - 0.87\, f_5(u(\eta + (X_{t-1}^{14} \times X_{t-2}^1))) \\
X_t^2 &\leftarrow \eta + 0.65\, f_1(u(\eta + X_{t-1}^2)) - 0.63\, f_3(u(\eta + X_{t-2}^2)) + 0.79\, f_3(u(\eta + X_{t-1}^5)) \\
X_t^3 &\leftarrow \eta - 0.76\, f_5(u(\eta + X_{t-1}^3)) - 0.59\, f_6(u(\eta + X_{t-1}^7)) - 0.85\, f_2(u(\eta + X_{t-1}^{15})) \\
&\quad -0.89\, f_5(u(\eta + (X_{t-2}^{13} \times X_{t-1}^7))) \\
X_t^4 &\leftarrow \eta - 0.7\, f_6(u(\eta + X_{t-1}^5)) - 0.86\, f_2(u(\eta + X_{t-2}^8)) + 0.53\, f_1(u(\eta + (X_{t-1}^4 \times X_{t-2}^9))) \\
X_t^5 &\leftarrow \eta + 0.54\, f_2(u(\eta + (X_{t-1}^{14} \times X_{t-2}^6))) \\
X_t^6 &\leftarrow \eta - 0.85\, f_2(u(\eta + X_{t-1}^6)) - 0.79\, f_3(u(\eta + X_{t-2}^3)) + 0.59\, f_1(u(\eta + X_{t-1}^4)) \\
&\quad +0.75\, f_3(u(\eta + X_t^1)) + 0.57\, f_2(u(\eta + X_{t-1}^{14})) \\
X_t^7 &\leftarrow \eta + 0.74\, f_1(u(\eta + X_{t-1}^7)) + 0.54\, f_6(u(\eta + X_{t-1}^9)) - 0.53\, f_2(u(\eta + (X_{t-1}^9 \times X_{t-1}^7))) \\
X_t^8 &\leftarrow \eta \times (-0.63\, f_1(u(\eta + X_{t-1}^6)) + 0.81\, f_5(u(\eta + X_t^{13})) + 0.53\, f_6(u(\eta + (X_{t-2}^6 \times X_{t-1}^6))) \\
&\quad -0.69\, f_6(u(\eta + (X_t^{13} \times X_{t-1}^6)))) \\
X_t^9 &\leftarrow \eta + 0.79\, f_3(u(\eta + X_{t-2}^4)) + 0.69\, f_6(u(\eta + (X_{t-1}^9 \times X_{t-1}^{15}))) \\
X_t^{10} &\leftarrow \eta + 0.54\, f_6(u(\eta + X_{t-1}^{10})) \\
X_t^{11} &\leftarrow \eta + 0.83\, f_6(u(\eta + X_{t-1}^{11})) - 0.76\, f_4(u(\eta + X_{t-1}^{13})) - 0.73\, f_3(u(\eta + X_{t-1}^2)) \\
&\quad +0.74\, f_2(u(\eta + X_t^4)) - 0.87\, f_2(u(\eta + X_{t-2}^{10})) + 0.72\, f_4(u(\eta + X_{t-1}^{12})) \\
&\quad -0.73\, f_1(u(\eta + (X_{t-2}^{10} \times X_{t-1}^{13}))) \\
X_t^{12} &\leftarrow \eta + 0.7\, f_3(u(\eta + X_{t-1}^{10})) - 0.55\, f_5(u(\eta + X_t^9)) - 0.54\, f_5(u(\eta + (X_{t-1}^{12} \times X_{t-1}^{10}))) \\
X_t^{13} &\leftarrow \eta - 0.62\, f_3(u(\eta + X_{t-2}^{14})) - 0.61\, f_1(u(\eta + (X_{t-1}^{13} \times X_{t-2}^{14}))) \\
X_t^{14} &\leftarrow \eta - 0.78\, f_6(u(\eta + X_{t-1}^{14})) \\
X_t^{15} &\leftarrow \eta - 0.68\, f_4(u(\eta + X_{t-1}^{15})) + 0.85\, f_4(u(\eta + X_{t-2}^{15})) - 0.6\, f_5(u(\eta + X_{t-2}^{10})) \\
&\quad +0.68\, f_6(u(\eta + X_{t-1}^{14})) + 0.81\, f_4(u(\eta + (X_{t-1}^{14} \times X_{t-2}^{10})))
\end{aligned}
$$

Fonctions

$$
\begin{aligned}
u(x) &= \max(-1, \min(1, x)) \\
f_1(x) &= x \\
f_2(x) &= x(1 - 4e^{-\frac{x^2}{2}})/1.52387 \\
f_3(x) &= 4x^2 \\
f_4(x) &= 8x^3 \\
f_5(x) &= 16x^4 \\
f_6(x) &= \cos(\pi x)
\end{aligned}
$$

Bruits

Les η sont des bruits blancs générés pour chaque nœud ou contribution en utilisant une distribution normale :

$$\eta \sim \mathcal{N}(0, 0.1)$$

Comme pour le modèle précédent, dix jeux de données ont été générés avec dix initialisations du générateur aléatoire différentes. Des reconstructions de réseaux ont été effectuées sur ces dix jeux de données avec tMIIC, PC et PCMCI+. Pour PC et PCMCI+, les trois noyaux ParCorr, GPDC et CMIknn ont été testés. Toutes les méthodes ont été lancées avec leurs paramètres par défaut à une exception : pour tMIIC, le nombre de threads à été réglé sur 8 pour obtenir une comparaison juste avec Tigramite dont les noyaux utilisent le processeur à 100 %.

Les méthodes/noyaux ont été testés sur un nombre progressif de pas de temps avec une limite à 30 minutes de temps d'exécution maximum pour maintenir la durée totale du benchmark acceptable.

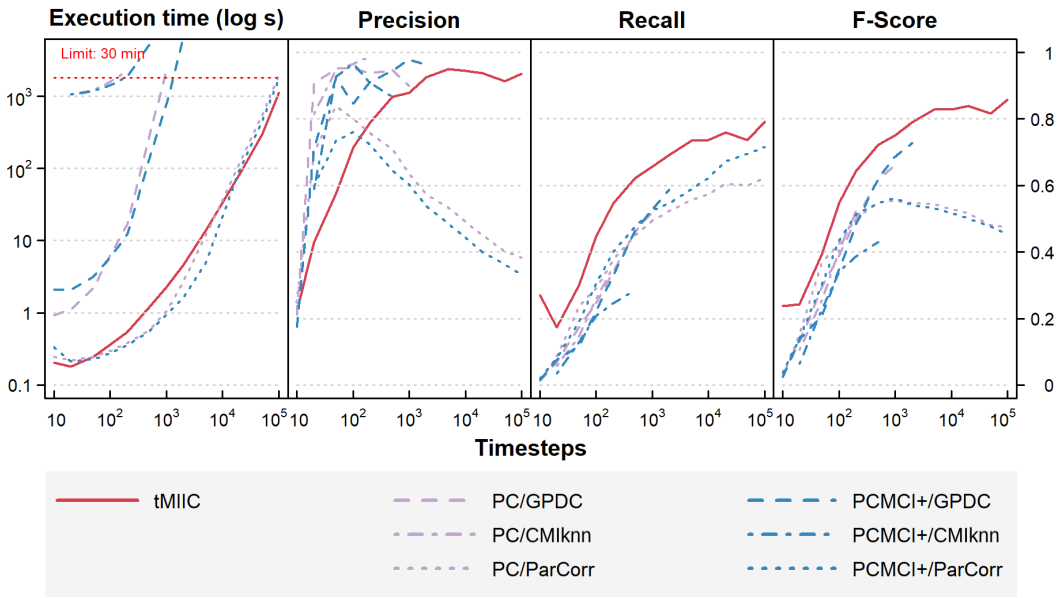


FIGURE 5.13 – **Evaluation par benchmark de la découverte causale tMIIC sur des séries temporelles contenant des combinaisons non-linéaires des contributions.**

Nous avons de nouveau effectué la moyenne des temps d'exécution et des scores (Précision, Rappel, F-score) obtenus sur les 10 jeux de données. Il ressort que pour les temps d'exécution, tMIIC est proche des noyaux les plus rapides et tMIIC surpasse à la fois les algorithmes PC et PCMCI+, quel que soit le noyau utilisé, sur le F-score orienté.

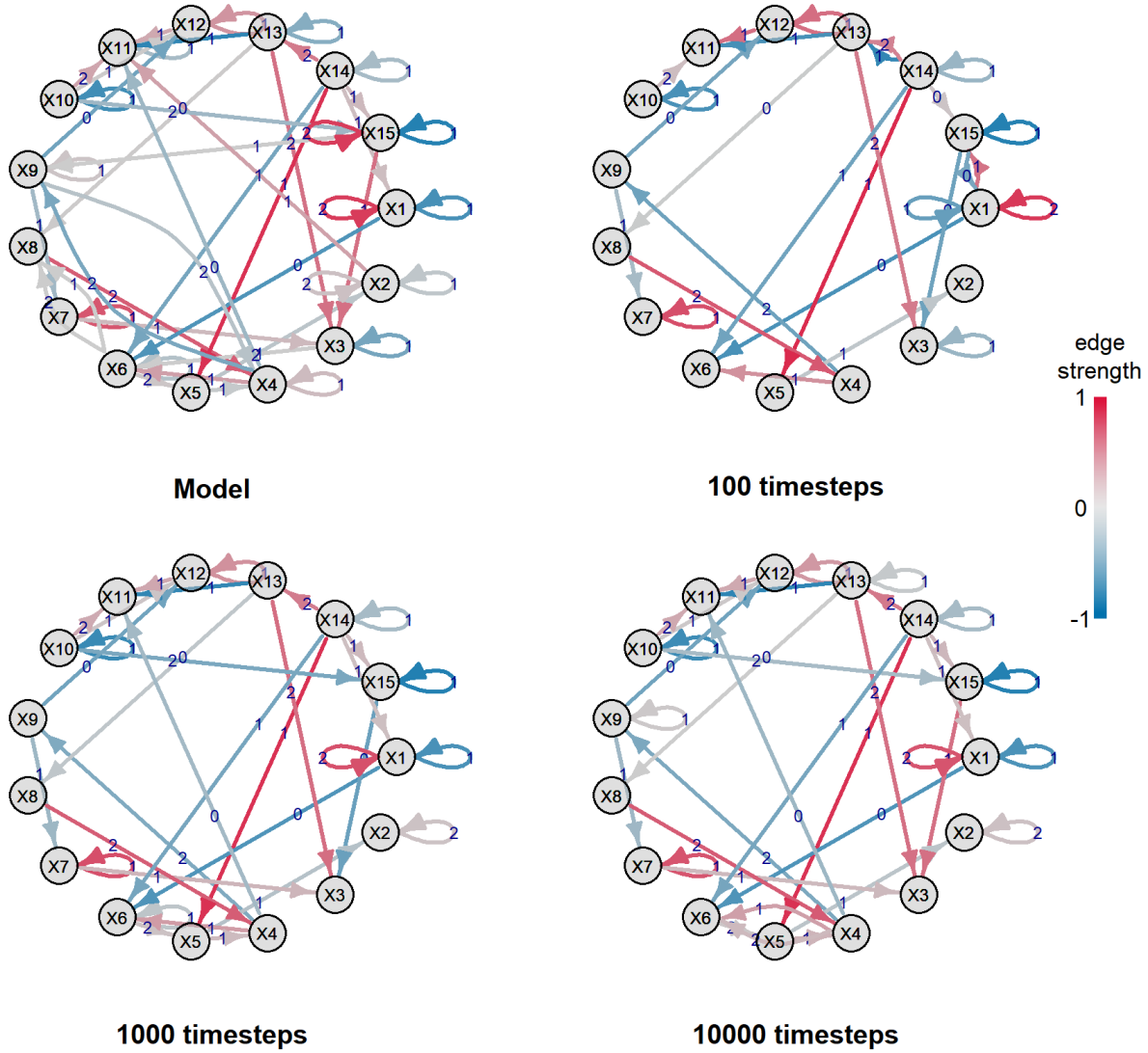


FIGURE 5.14 – **Exemples de reconstructions de réseaux causaux par tMIIC sur des séries temporelles générées.** A partir de 1 000 pas de temps, nous obtenons une reconstruction proche du modèle. "Edge strength" est le coefficient de corrélation de Spearman, éventuellement partiel s'il y a des contributeurs.

5.3.2 Découverte sur une fenêtre trop large

Notre objectif étant, in fine, d'appliquer tMIIC sur des données de la vie réelle, pour lesquelles la dynamique n'est pas connue, nous avons évalué sur les bases de ces benchmarks, l'impact causé par l'utilisation d'une fenêtre de découverte trop importante.

Pour cela, nous avons réutilisé les jeux d'essais du premier modèle en augmentant progressivement la taille de la fenêtre de la découverte causale que nous avons augmentée à partir de 2, puis 5 et pour finir 10 pas de temps.

Les résultats, présentés sur la figure 5.15, s'ils montrent une légère baisse du score qui est normale, nous permettent également d'observer que les graphes obtenus restent globalement invariants malgré l'augmentation de la fenêtre de recherche au-delà de la taille nécessaire pour couvrir la dynamique du système.

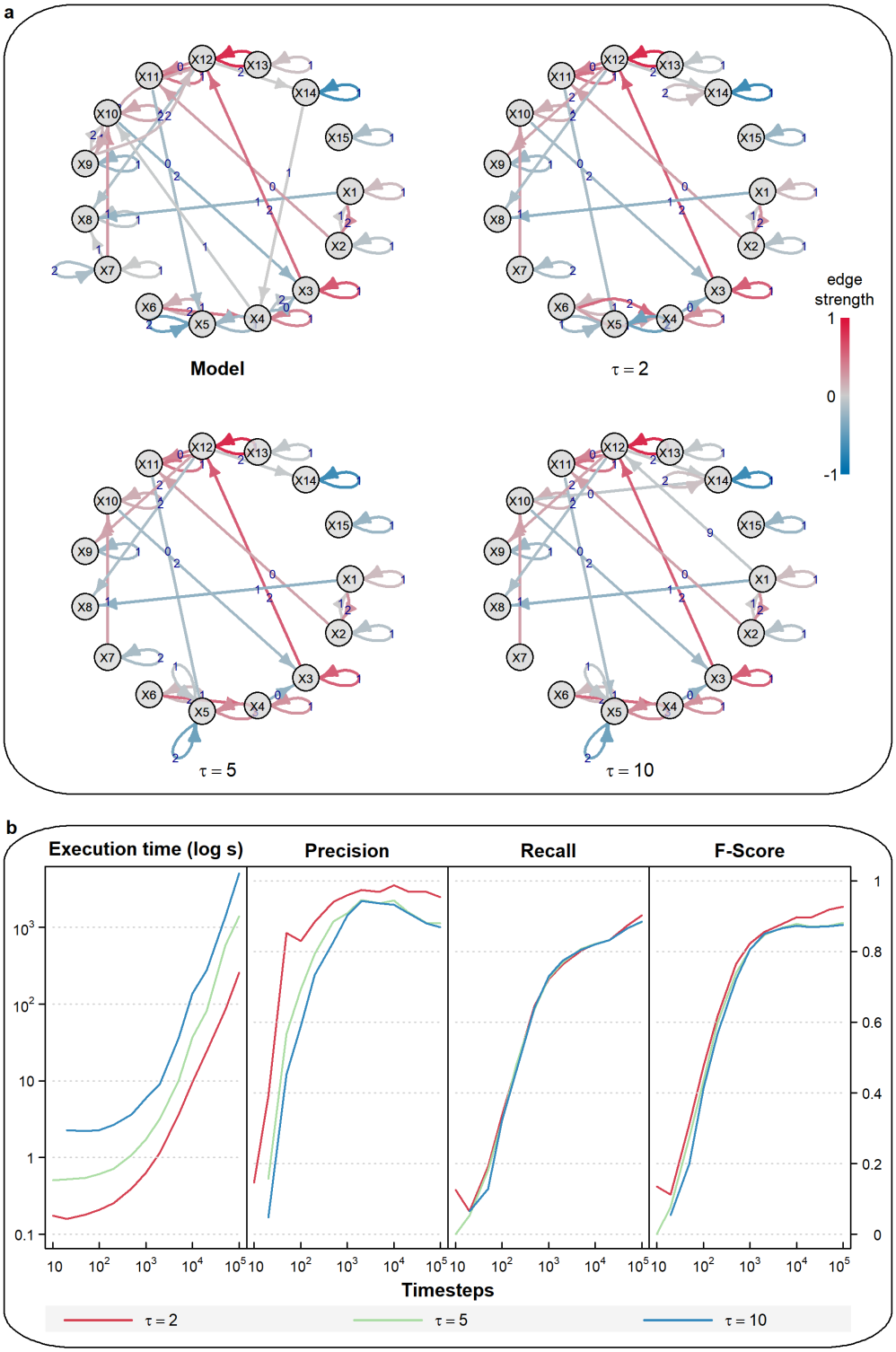


FIGURE 5.15 – **Insensibilité de tMIIC à une fenêtre de découverte τ surestimée. a**, Exemple de modèle de réseau causal temporel avec une dynamique maximum $\tau = 2$. Réseaux causaux temporels correspondants déduits par tMIIC, à partir de séries temporelles stationnaires de 1 000 pas de temps, en supposant différentes fenêtres de découverte $\tau = 2$, 5 ou 10. "Edge strength" est le coefficient de corrélation de Spearman, éventuellement partiel s'il y a des contributeurs. **b**, Temps d'exécution et scores (Précision, Rappel, Fscore) des reconstructions du réseau causal temporel par tMIIC pour $\tau = 2$, 5 ou 10, en moyenne sur les dix séries temporelles stationnaires de 10 à 10^5 pas de temps. La surestimation de la fenêtre de découverte a peu d'impact sur les réseaux reconstruits, tant que les séries temporelles sont stationnaires.

5.3.3 Impact de la non-stationnarité

Dans la mesure où nous avons posé l'hypothèse de stationnarité des variables pour tMIIC, nous avons ensuite souhaité évaluer l'impact que pouvait avoir la non-stationnarité d'une variable sur notre réseau. Pour cela, nous avons introduit dans le modèle simple une tendance et une saisonnalité sur, respectivement, les nœuds $X13$ et $X8$.

Comme pour le benchmark précédent, nous avons ensuite procédé à des reconstructions causales en augmentant progressivement la taille de la fenêtre de 2, 5 puis 10 pas de temps dans le passé.

Les réseaux obtenus regroupés sur la figure 5.16, révèlent de façon très intéressante l'apparition de multiples boucles sur les variables non stationnaires, qui traduisent que tMIIC a besoin d'effets de plus en plus lointains pour arriver à expliquer leur dynamique non stationnaire. Par contre, il apparait également sur ce benchmark, qu'à l'exclusion de ces boucles, le reste du graphe est globalement inchangé.

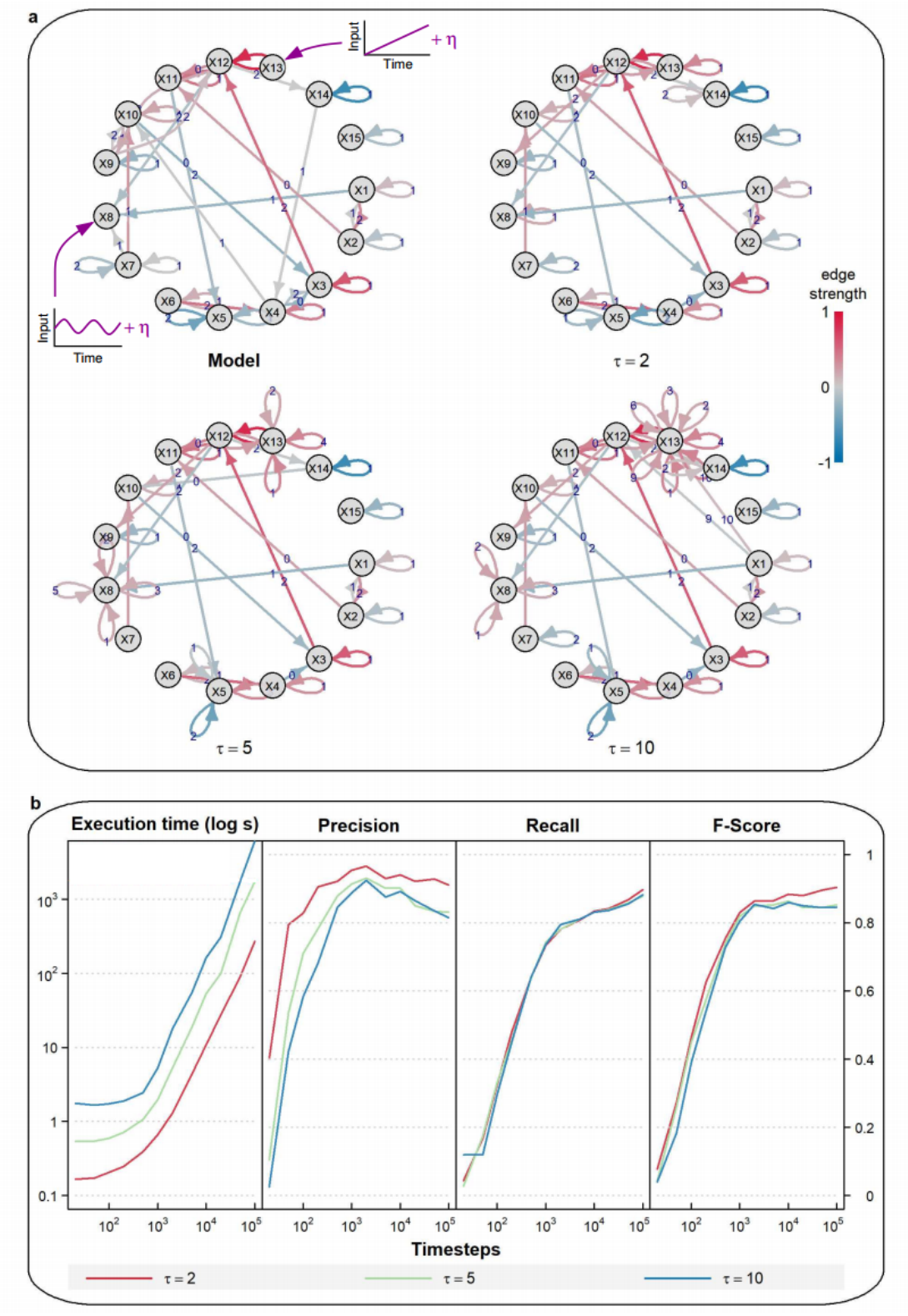


FIGURE 5.16 – **Sensibilité de tMIIC aux variables non stationnaires.** a, Exemple de modèle de réseau causal temporel ($\tau = 2$) avec une entrée périodique basse fréquence ($T = 100$) appliquée à $X8$ et une tendance linéaire temporelle appliquée à $X13$. Réseaux causaux temporels correspondants déduits par tMIIC à partir de 1 000 pas de temps, y compris pour les variables non stationnaires de $X8$ et $X13$. "Edge strength" est le coefficient de corrélation de Spearman, éventuellement partiel s'il y a des contributeurs. b, Temps d'exécution et scores (Précision, Rappel, Fscore ignorant les boucles sur $X8$ et $X13$) des reconstructions de réseaux causaux par tMIIC pour $\tau = 2$, 5 ou 10, en moyenne sur dix séries temporelles de 10 à 10^5 pas de temps.

5.4 Application

L'application a été réalisée sur des séquences d'images provenant de cultures cellulaires filmées au microscope (une image toutes les 2 minutes). Ces séquences d'images chronologiques de cellules vivantes proviennent d'une étude de preuve de concept qui démontre les effets d'un médicament anticancéreux (les anticorps monoclonaux trastuzumab, nom de marque Herceptin, utilisés pour traiter les cancers du sein HER2+) sur un micro-environnement tumoral reconstitué comprenant des cellules cancéreuses, des cellules immunitaires, des fibroblastes (CAF) et des cellules endothéliales [55].

Si l'objectif de M. Nguyen avec ces expériences était de démontrer l'intérêt des micro-environnements tumoraux reconstitués pour les études pré-cliniques de médicaments, pour notre projet de découverte causale, l'intérêt était de disposer d'un grand nombre d'images (46 935) ordonnées dans le temps et d'évaluer si tMIIC était en mesure, uniquement à partir des données, de retrouver les relations causales connues voire d'en identifier de nouvelles.

5.4.1 Préparation des données

De ces images, ont été extraites les caractéristiques cellulaires telles que la géométrie cellulaire, la vitesse, les interactions transitoires cellule-cellule et les contacts persistants à l'aide de l'outil CellHunter+ [55].

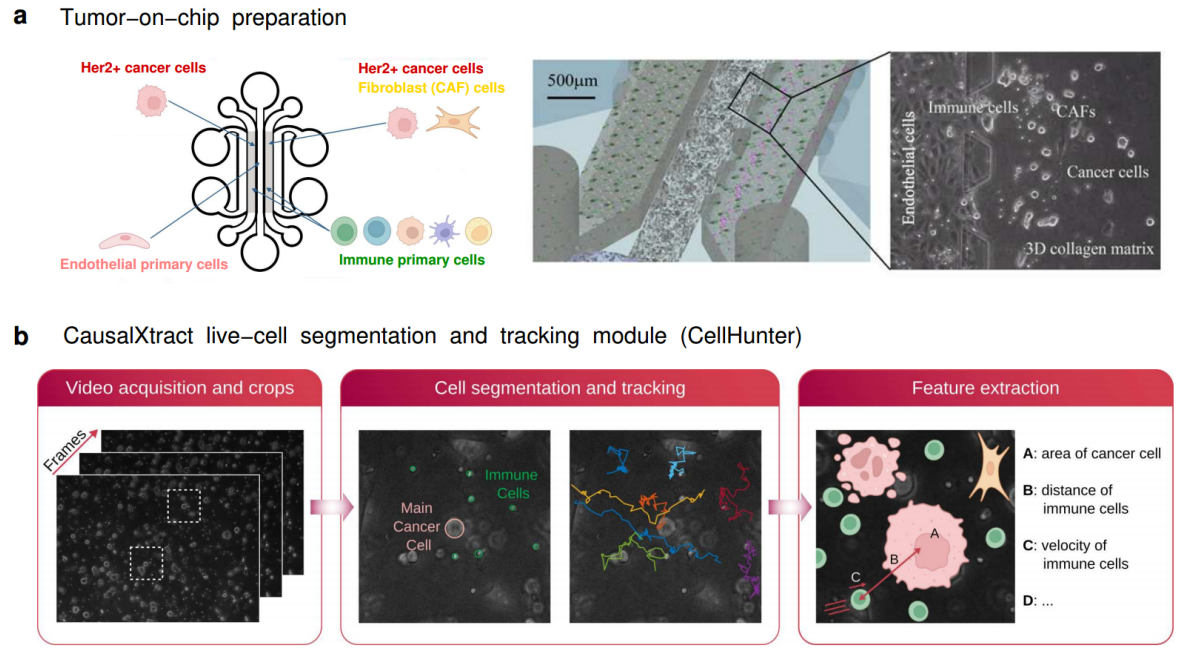


FIGURE 5.17 – **Données d'application de tMIIC a**, Ecosystèmes tumoraux à cellules vivantes reconstitués ex vivo à l'aide de la technologie tumor-on-chip. **b**, segmentation et suivi des cellules vivantes basés sur le logiciel CellHunter+

Le fichier obtenu en utilisant CellHunter+ se présente sous la forme suivante :

	IDExp	IDCell	IDFrame	CAFsPresence	Treatment
1	id20161230_pos1_roi6	1	1	0	0
2	id20161230_pos1_roi6	1	2	0	0
3	id20161230_pos1_roi6	1	3	0	0
4	id20161230_pos1_roi6	1	4	0	0
5	id20161230_pos1_roi6	1	5	0	0
6	...	...	...	...	...

	Loc_Apoptosis	Loc_Duplication	Area	InstSpeed	InstChangeShape
1	0	0	1643	NA	NA
2	0	0	1659	0.3335473	0.2221635
3	0	0	1572	1.2446499	1.2213221
4	0	0	1404	0.6662009	2.458137
5	0	0	1509	1.3127857	1.5524931
6	...	...	...	...	...

	Perimeter	Eccentricity	Circularity	TotDist	NetDist
1	176.197	0.5183651	0.665044	NA	NA
2	185.642	0.5099649	0.6049281	0.3335473	0.3335473
3	187.949	0.5143502	0.5592196	1.5781972	1.3723811
4	227.572	0.36511	0.3406743	2.2443981	1.9055449
5	215.813	0.5676372	0.4071401	3.5571838	1.4888839
6	...	...	...	...	...

	NumCellsBehind_r2	NumCellsInFront_r2	sum_cum_int_r2	mean_v_r2
1	0	0	NA	NA
2	0	0	NA	NA
3	0	0	NA	NA
4	0	0	NA	NA
5	0	0	NA	NA
6	...	...	...	...

	mean_v_r1	count_cell_at_frame_r1	Glob_Apoptosis	IDint
1	NA	NA	0	0
2	NA	NA	0	0
3	NA	NA	0	0
4	NA	NA	0	0
5	NA	NA	0	0
6	...	...	...	...

Positionnement du temps en première colonne

Dans l'ensemble de données, nous pouvons voir plusieurs caractéristiques d'identification dans les trois premières colonnes : $IDExp$, $IDCell$, $IDFrame$. Ces informations peuvent être intéressantes pour d'autres analyses mais seules l'information de pas de temps est utile pour la découverte causale avec tMIIC. Si nous avions conservé $IDExp$ et $IDCell$, tMIIC aurait essayé de trouver des relations causales entre ces variables et les autres caractéristiques du jeu de données, nous avons donc supprimé ces deux co-

lonnes. Nous nous sommes ensuite assuré que *IDFrame*, l'information de temps était dans la première colonne, puisque tMIIC l'attend à cette position.

Le jeu de données modifié se présente comme suit :

	IDFrame	CAFsPresence	Treatment	Loc_Apoptosis	Loc_Duplication	...
1	1	0	0	0	0	...
2	2	0	0	0	0	...
3	3	0	0	0	0	...
4	4	0	0	0	0	...
5	5	0	0	0	0	...
6	...	...	...	...	...	...

À ce stade, tMIIC pouvait déjà être appliqué à l'ensemble de données, ce que nous avons d'ailleurs fait car nous étions impatients de voir les premiers résultats de tMIIC, cependant, les biais présents dans le jeu de données affectaient les résultats et une inspection des données a été nécessaire pour nous assurer de leur qualité, ainsi qu'un ajustement des paramètres de tMIIC pour obtenir un meilleur résultat.

Vérifications supplémentaires sur les informations de temps

Lorsque le tMIIC traite le jeu de données d'entrée, il considère la première colonne comme représentant les pas de temps. Ces informations de pas de temps sont utilisées pour identifier les trajectoires (chaque trajectoire est une réalisation de la même expérience ou processus) et une nouvelle trajectoire est détectée lorsque le pas de temps de la ligne précédente est supérieur au pas de temps de la ligne suivante.

Nous nous sommes donc assurés que :

- Nous obtenions bien le nombre de trajectoires attendu (36)
- Toutes les trajectoires commençaient par un pas de temps de 1
- il n'y avait pas d'écarts entre les pas de temps de lignes consécutives
- il y n'y avait pas de valeurs de pas de temps inchangées entre des lignes consécutives

Sur ces contrôles, le seul point que nous avons remarqué est que certaines trajectoires ne commençaient pas par le pas de temps 1 (donc pas à la première image de la série de clichés), et les pas de temps de ces trajectoires ont donc été renumérotés.

Exemple de correction des pas de temps effectuée :

IDFrame	Perimeter	Eccentricity
4	176.197	0.5183651
5	185.642	0.5099649
6	187.949	0.5143502
7	227.572	0.36511
8	225.165	0.43525
9	215.813	0.5676372

IDFrame	Perimeter	Eccentricity
1	176.197	0.5183651
2	185.642	0.5099649
3	187.949	0.5143502
4	227.572	0.36511
5	225.165	0.43525
6	215.813	0.5676372

L'expérience acquise lors du contrôle des pas de temps a permis d'améliorer tMIIC puisque l'ensemble de ces vérifications a été intégré dans tMIIC et des avertissements sont affichés en cas de problème potentiel. La re-numérotation à partir de 1 des pas de temps a elle aussi été intégrée de manière automatique dans tMIIC et n'est plus à faire manuellement par l'utilisateur.

Inspection des données

Suite à la vérification des pas de temps, nous avons procédé à une inspection des données, deux trajectoires sont tracées sur les pages suivantes à titre d'exemples. A partir de ces graphiques, nous avons pu contrôler la dynamique par rapport à nos connaissances sur les variables et identifier certaines caractéristiques de ces variables.

Quelques caractéristiques intéressantes à noter :

- variables contextuelles : les conditions expérimentales, ici *Treatment* et *CAFs-Presence* (CAF : fibroblaste associé au cancer)
- variables discrètes : ici, encore *Treatment* et *CAFsPresence*, mais aussi *Loc_Apoptosis*, *Loc_Duplication*, *Glob_Apoptosis* et *IDint*
- variables continues : *Area*, *Eccentricity*, ..., toutes les variables à virgule flottante
- variables non stationnaires : à l'évidence *Totdist*, *max_cum_int_r2* et potentiellement *Eccentricity* ou *StraightIndex*.
- variables avec un nombre élevé de valeurs manquantes (NA) : une majorité des caractéristiques mesurant les interactions cancer-cellules immunitaires (les variables se terminant par *r1* ou *r2*)
- variables fortement corrélées : ici, les variables *count_cell_at_frame_r2*, *NumCellsInFront_r2*, *NumCellsBehind_r2* et *IDint* semblent être fortement liées. Ces variables posent la question de savoir s'il agit de la même information ou presque ?
- d'autres points à approfondir peuvent aussi être notés : par exemple, pourquoi avons-nous deux variables concernant l'apoptose ? Quelle est la différence entre *Loc_Apoptosis* et *Glob_Apoptosis* ? Autre exemple, quelle est la différence entre *NumCellsBehind_r2* et *NumCellsInFront_r2* ?

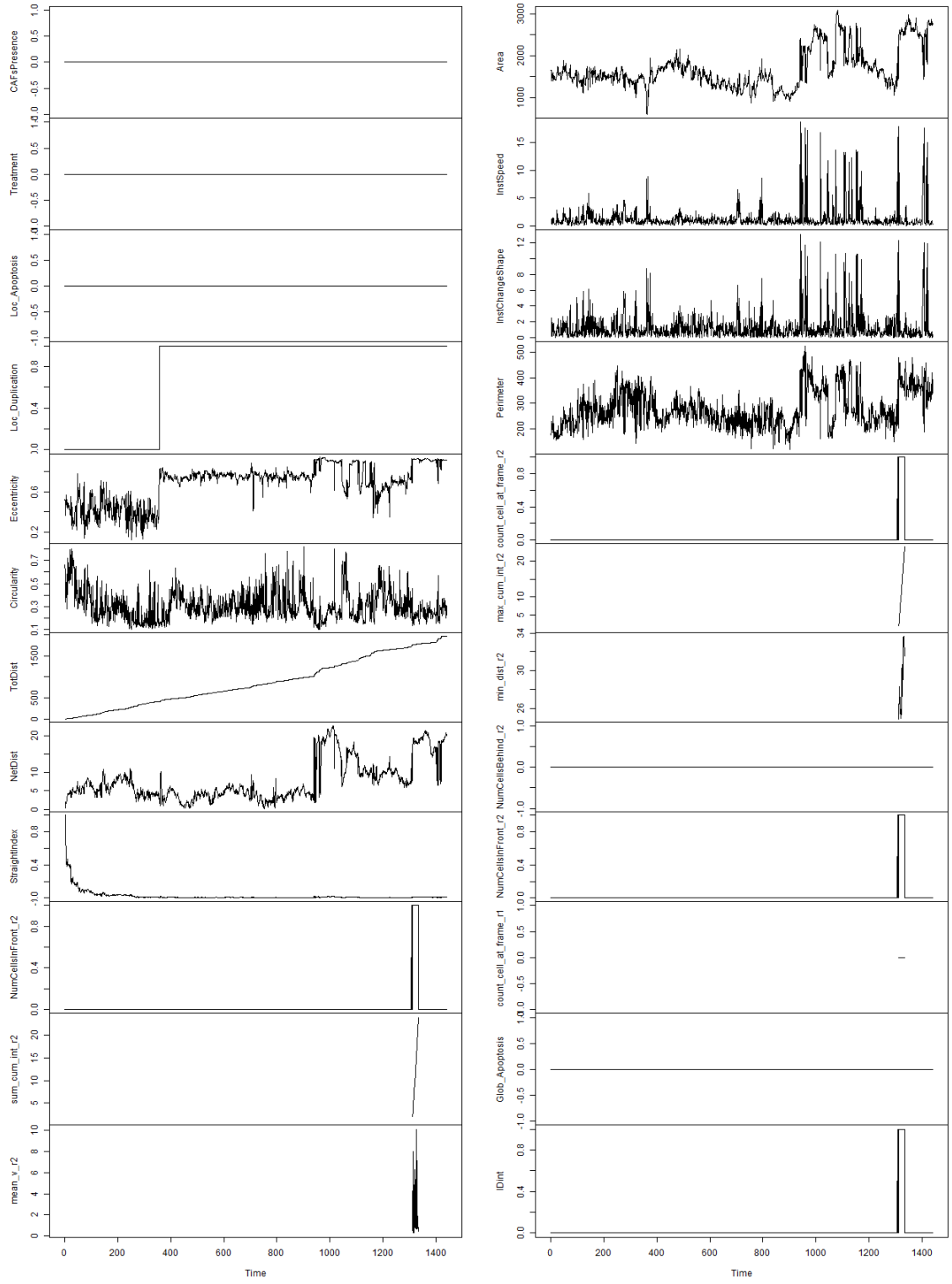


FIGURE 5.18 – **Exemple de trajectoire.** 1^{re} trajectoire sans traitement, ni fibroblaste

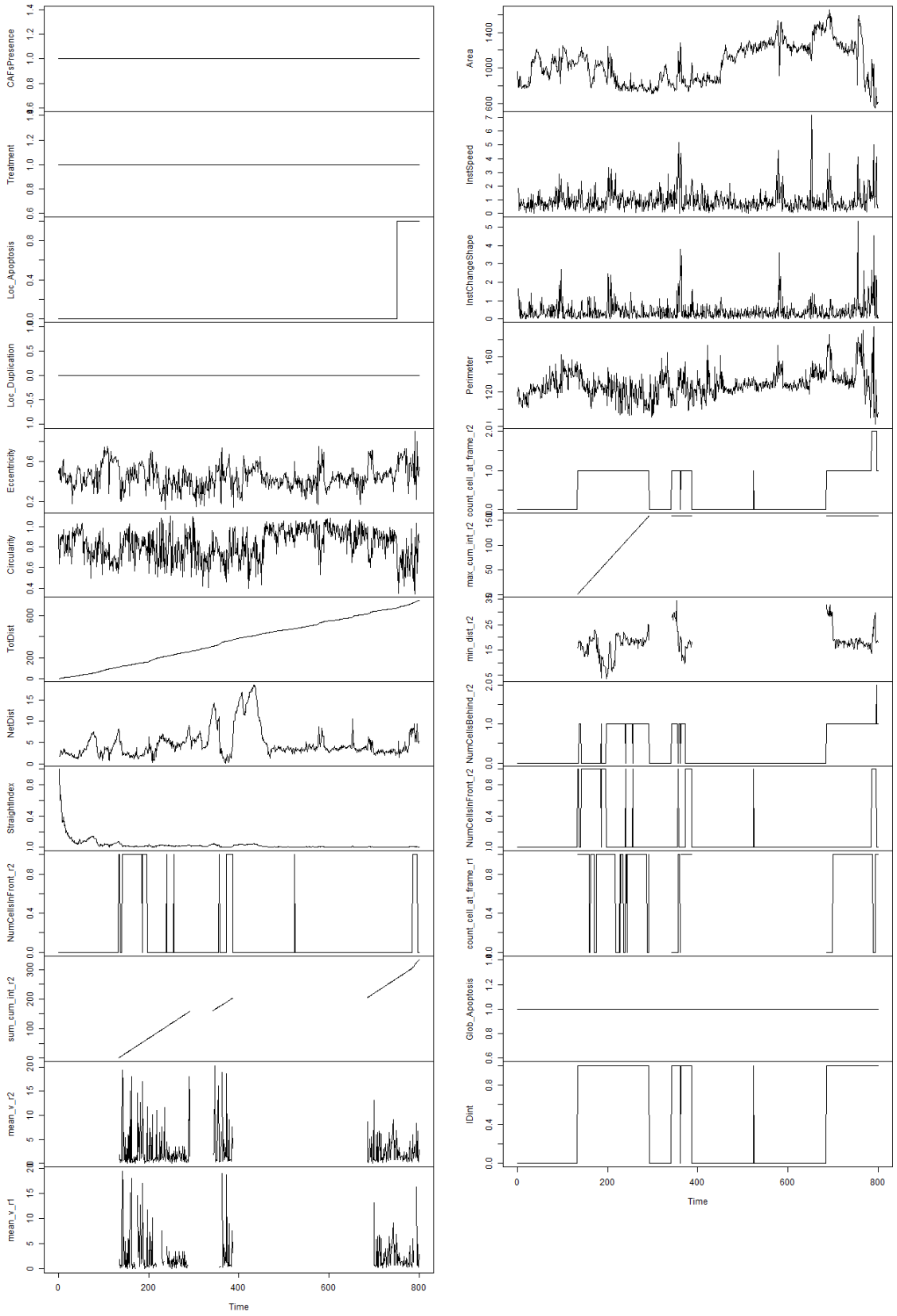


FIGURE 5.19 – **Exemple de trajectoire.** 27^{me} trajectoire avec traitement et fibroblastes

Les variables que nous avons notées lors de l'inspection visuelle vont être étudiées et prises en compte pour améliorer le résultat de tMIIC.

Variables non stationnaires

Comme tMIIC repose sur l'hypothèse que la dynamique des données est stationnaire, toutes les variables susceptibles de violer cette hypothèse ont été revues. Lors de l'inspection visuelle, nous mentionnions dans notre exemple $TotDist$, $max\_cum\_int\_r2$ et potentiellement $Eccentricity$ et $StraightIndex$ comme non stationnaires.

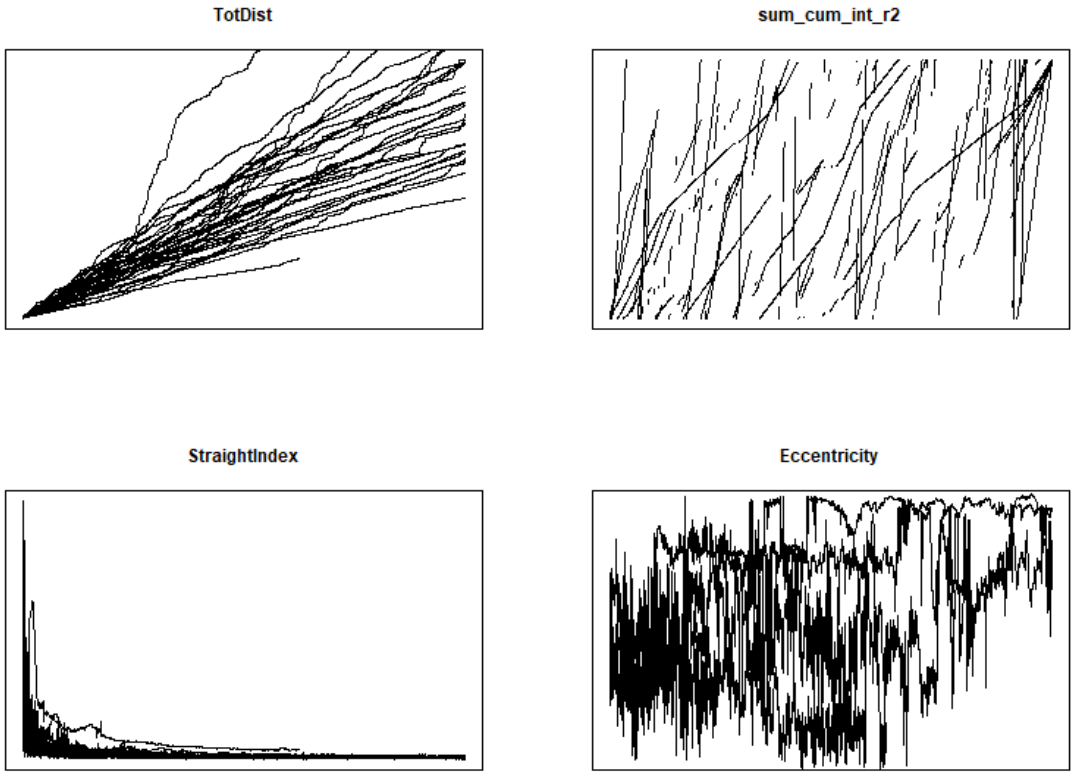


FIGURE 5.20 – Examen des caractéristiques non stationnaires.

À partir de ces graphiques, nous pouvons immédiatement voir que certaines variables ne sont clairement pas stationnaires. $TotDist$ et $max\_cum\_int\_r2$ peuvent facilement être identifiées comme étant cumulatives et doivent être supprimées. Il en va de même pour $StraightIndex$ qui disparaît avec le temps et a été retirée. De ces quatre variables, nous n'avons retenu que $Eccentricity$ car, même si elle n'est pas complètement stationnaire, elle n'est ni cumulative, ni s'évanouissant au fil du temps.

Une alternative à la suppression des variables aurait pu être de les rendre stationnaires, c'est-à-dire en utilisant par exemple la dérivée de ces variables. Dans notre application, nous n'avons pas retenu cette méthode car pour les variables $TotDist$ et $max\_cum\_int\_r2$, nous avons déjà dans le jeu de données, respectivement, $NetDist$ et $count\_cell\_at\_frame\_r2$ qui correspondent à la dérivée des variables cumulatives.

Violation de la temporalité

Lors de l'inspection visuelle, nous avions repéré deux caractéristiques relatives à l'apoptose : $Loc\_Apoptosis$ et $Glob\_Apoptosis$. Un examen plus approfondi de ces deux

variables montre que *Glob_Apoptosis* sait dès le départ (donc à l'avance du point de vue temporel) si la cellule va mourir au cours de l'expérience.

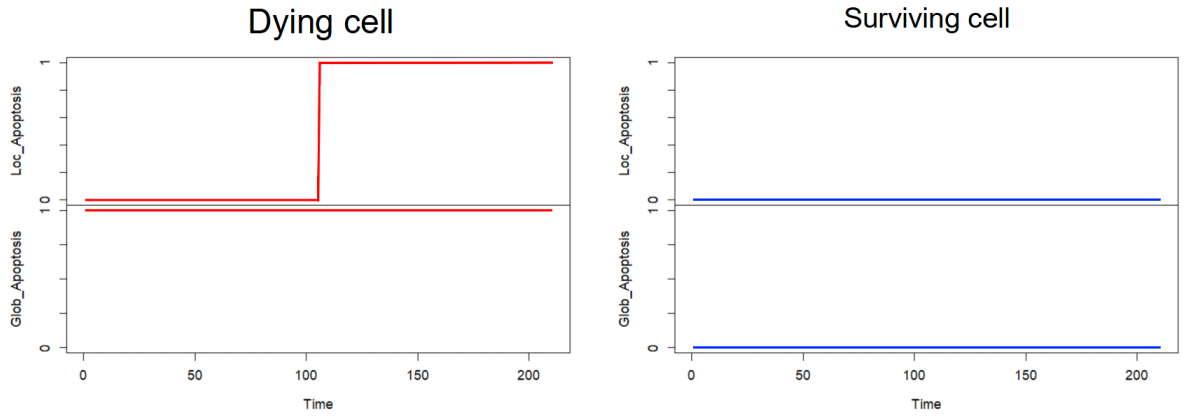


FIGURE 5.21 – **Focus sur la violation de la temporalité.**

Comme *Glob_Apoptosis* sait dès le début quel sera le sort de la cellule, cette variable devait être supprimée car elle est en contradiction avec l'hypothèse de temporalité selon laquelle le futur ne peut influencer ni le passé, ni le présent.

Variables redondantes

Comme nous avions remarqué que $count\_cell\_at\_frame\_r2$ et $IDint$ semblent être fortement liées, nous avons examiné de plus près ces deux variables. Il s'est avéré que $IDint$ est simplement un proxy de $count\_cell\_at\_frame\_r2$: $IDint$ vaut 1 lorsque $count\_cell\_at\_frame\_r2$ est > 0 et IDint $= 0$ lorsque $count\_cell\_at\_frame\_r2 = 0$.

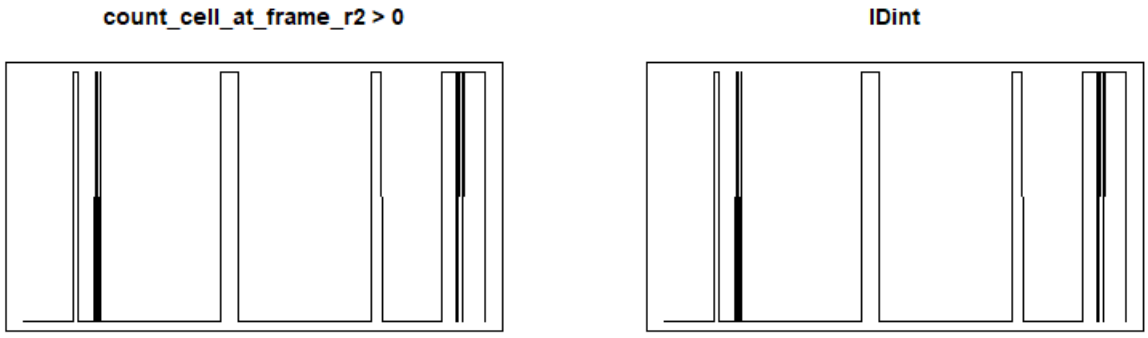


FIGURE 5.22 – **Variables proxy** : $IDint$ est un proxy pour $count\_cell\_at\_frame\_r2$.

L'inclusion de variables proxy pour une découverte causale peut conduire à des résultats erronés, car lors de l'évaluation d'une arête où un nœud est l'une de ces variables redondantes, le conditionnement sur l'autre variable redondante conduira à une information nulle ou très faible, de sorte que l'arête est donc susceptible d'être supprimée du graphe même s'il existe réellement de l'information sur cette arête. Par conséquent, $IDint$ a été supprimée car nous avons conservé la variable la plus informative $count\_cell\_at\_frame\_r2$.

Variables non pertinentes

Il y avait également, lors de l'inspection visuelle, des questions relatives à la signification de $NumCellsInFront\_r2$ et $NumCellsBehind\_r2$. Après discussion avec l'équipe de CellHunter+, il est apparu que ces deux caractéristiques sont relatives à la position où les cellules immunitaires sont détectées dans l'image et qu'il n'y a aucune signification biologique par rapport à l'arrière ou l'avant.

De plus, ces deux variables sont étroitement liées à $count\_cell\_at\_frame\_r2$ car $count\_cell\_at\_frame\_r2 = NumCellsBehind\_r2 + NumCellsInFront\_r2$. Nous avons donc décidé d'écarter ces deux variables : non pertinentes pour nos objectifs et trop similaires avec $count\_cell\_at\_frame\_r2$.

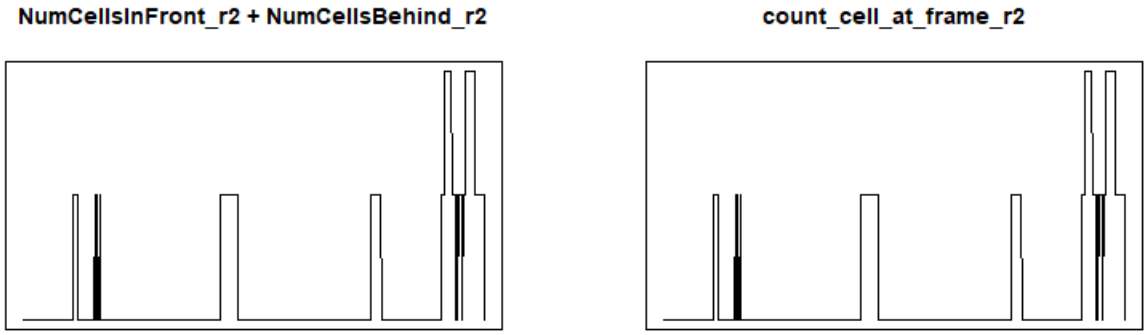


FIGURE 5.23 – **Variables non pertinentes** : $NumCellsInFront\_r2$ et $NumCellsBehind\_r2$ n'ont aucune signification biologique et leurs informations sont déjà présentes si l'on considère la variable $count\_cell\_at\_frame\_r2$.

Valeurs manquantes

Quoique tMIIC soit capable de gérer les valeurs manquantes, nous avons essayé de voir avec l'équipe de CellHunter+ s'il existait des moyens de réduire ces valeurs manquantes, très nombreuses sur quelques variables comme le montre la figure 5.24.

Ceci a permis d'identifier qu'un ensemble de NA pouvait en fait être affecté avec une valeur, particulièrement lorsqu'aucune cellule n'est présente dans le rayon $r2$ autour de la cellule cancéreuse (Fig 5.25).

L'ensemble de ces contrôles et modifications s'est révélé très utile pour la mise en place de CaulsalXtract (cf chapitre 6) car il nous a permis de modifier CellHunter+ pour constituer un pipeline complet d'analyse de dynamique cellulaire jusqu'à la reconstitution du réseau causal sous-jacent.

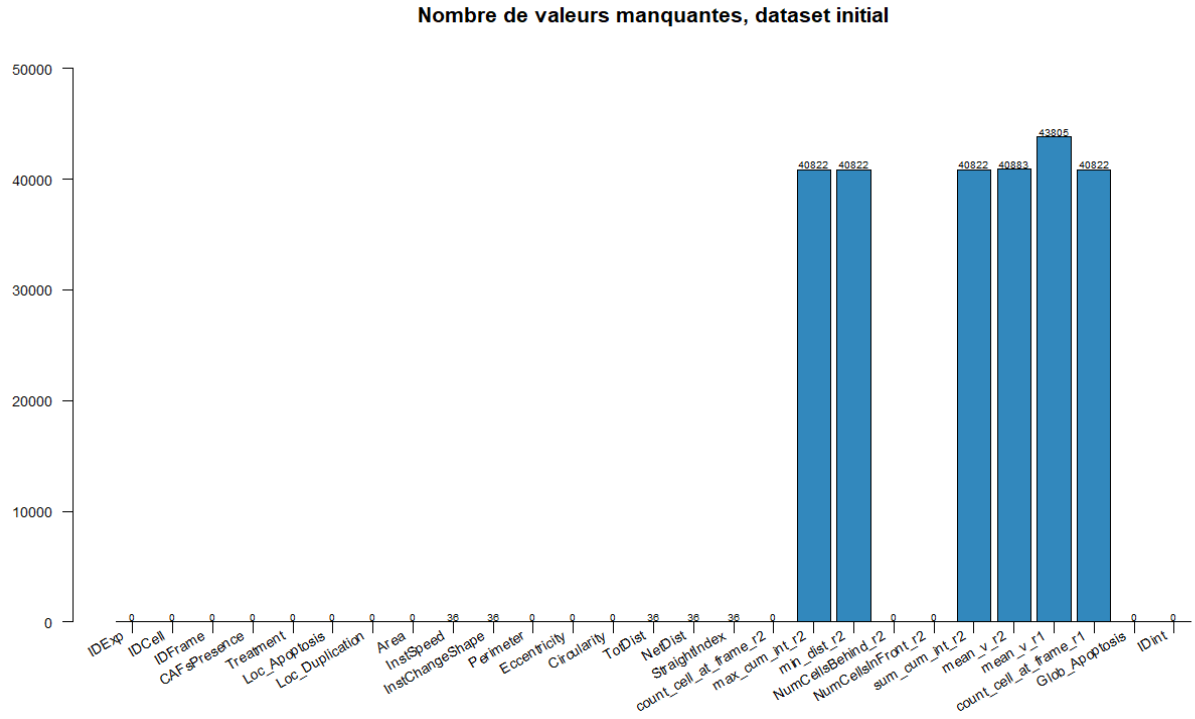


FIGURE 5.24 – Valeurs manquantes dans le jeu de données initial.

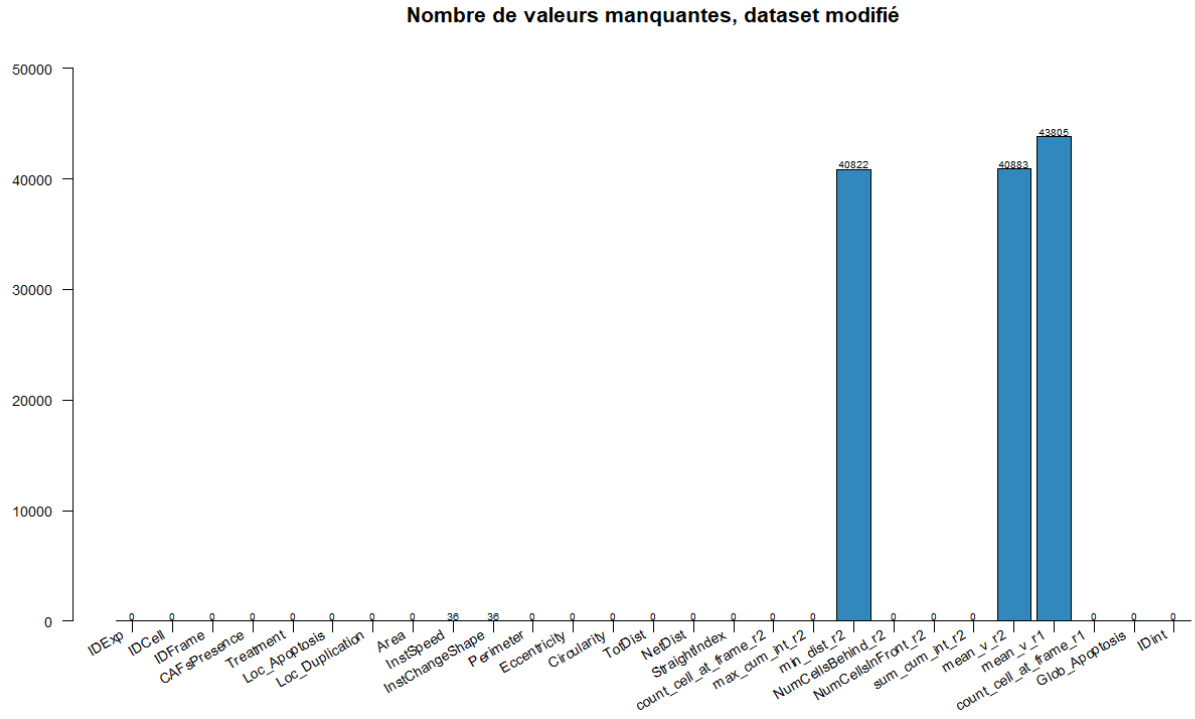


FIGURE 5.25 – Valeurs manquantes après modification.

5.4.2 Le fichier de paramétrage "state order"

Le *state order* est un paramètre supplémentaire de tMIIC et permet de spécifier des informations complémentaires sur les variables et la façon de les traiter.

Le format d'un *state order* est un tableau avec une colonne *var_names* qui permet d'identifier la variable et d'un ensemble de colonnes qui définissent des attributs spécifiques associés à cette variable. Je ne mentionnerai ici que les attributs *is_contextual* et *levels_increasing_order* que nous avons utilisés dans notre application bien que le champ d'application d'un *state order* soit plus large.

Niveaux de variables discrètes

Dans le cadre des variables discrètes, il est possible de spécifier via l'attribut *levels_increasing_order* un ordre croissant associé aux valeurs. Cet ordre est ensuite utilisé lors du post-traitement pour calculer le signe des arêtes en utilisant la corrélation de rang de Spearman. En plus de permettre le calcul de la corrélation des rangs de Spearman, cela permet d'améliorer le rendu visuel du graphe en colorant les effets positifs et négatifs.

Dans notre application, nous avions noté lors de l'inpection visuelle que les variables $Treatment$, $CAFsPresence$, $Loc\_Apoptosis$ et $Loc\_Duplication$ étaient logiques et elles ont donc été associées avec un attribut *levels_increasing_order* de $0, 1$.

Variables contextuelles

Comme remarqué également lors de l'inspection visuelle, nous avons deux conditions expérimentales $CAFsPresence$ et $Treatment$. Ces deux caractéristiques sont constantes pour chaque expérience et ne peuvent être la conséquence d'aucune autre caractéristique du jeu de données, car nous contrôlons ces conditions expérimentales.

Pour indiquer à tMIIC que ces variables ne peuvent être que des causes, nous avons donc utilisé l'attribut *is_contextual* qui a été défini 1 pour ces deux variables, et 0 pour toutes les autres.

Fichier de paramétrage obtenu

Deux attributs purement visuels *group* et *group_color* ont également été utilisés pour améliorer le rendu du graphe sur le serveur internet de MIIC. Pour notre application, nous avons défini trois groupes : conditions expérimentales, cellules cancéreuses et cellules immunitaires avec les couleurs associées à chaque groupe. Au final, notre *state order* a donc la forme suivante :

var_names	levels_ increasing _order	is_ contextual	group	group_ color
CAF_presence	0,1	1	experimental conditions	FF6600
treatment	0,1	1	experimental conditions	FF6600
apoptosis	0,1	0	cancer cells	D0D0D0
division	0,1	0	cancer cells	D0D0D0
area		0	cancer cells	D0D0D0
cancer_velocityy		0	cancer cells	D0D0D0
shape_change		0	cancer cells	D0D0D0
perimeter		0	cancer cells	D0D0D0
eccentricity		0	cancer cells	D0D0D0
circularity		0	cancer cells	D0D0D0
net_displacement		0	cancer cells	D0D0D0
directionality		0	cancer cells	D0D0D0
nb_interactions_r2		0	immune cells	00FF21
minimal_distance_r2		0	immune cells	00FF21
velocity_r2		0	immune cells	00FF21
velocity_r1		0	immune cells	00FF21
nb_interactions_r1		0	immune cells	00FF21

5.4.3 Détermination de la fenêtre de découverte

Une fois les données d'entrée vérifiées, nous avons recherché le nombre de pas de temps nécessaires pour couvrir l'ensemble de la dynamique lors de la découverte causale.

A ce stade, de nombreuses approches peuvent être utilisées pour déterminer ces paramètres. La première méthode que nous avons utilisée consistait à calculer et tracer des autocorrélations et corrélations croisées entre les variables, comme le montre la Fig.5.26, mais cette approche, plutôt visuelle, ne s'est pas révélée adaptée à notre application et s'applique mieux lorsque le nombre de variables est limité et avec une dynamique concise.

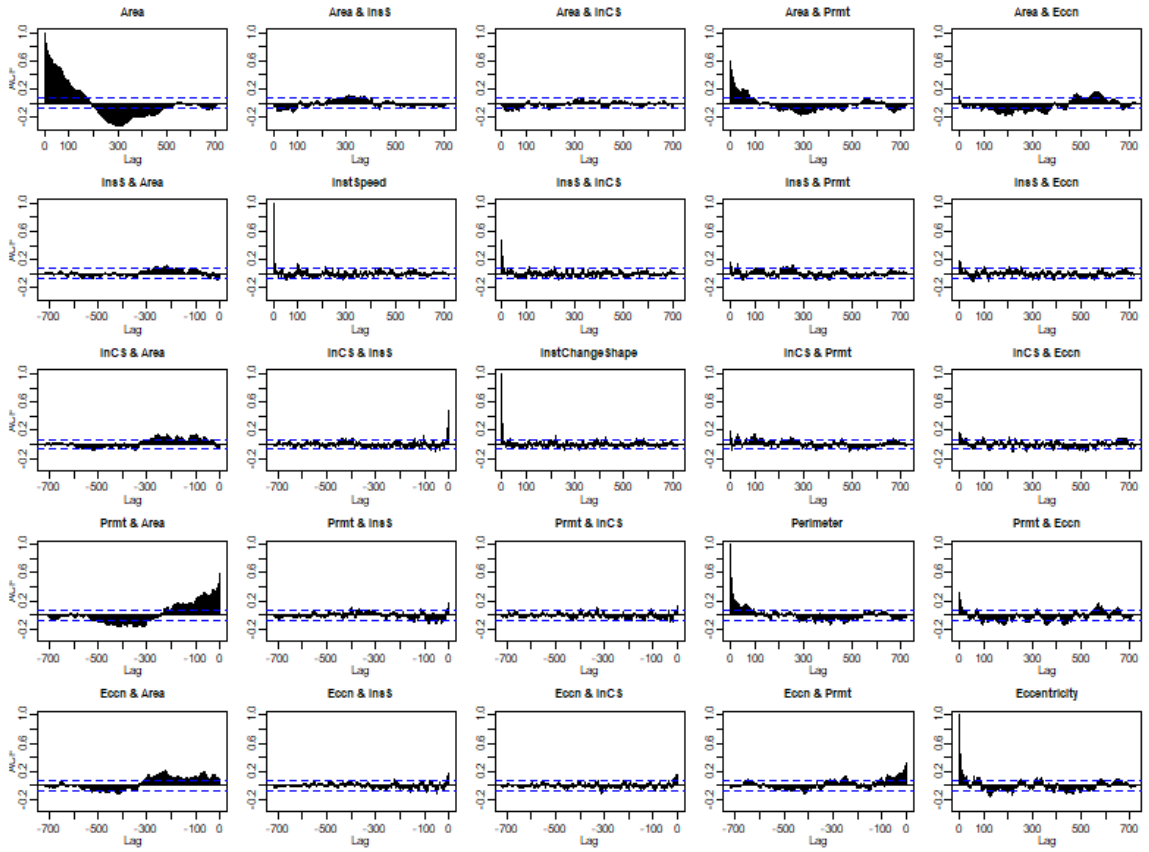


FIGURE 5.26 – **Graphiques de corrélations.** La dynamique semble passer de 20 pas de temps pour les variables les plus rapides (*instantaneous_cancer_velocity* et *instantaneous_shape_change*), puis 100-150 pas de temps (*area* et *perimeter*) et jusqu'à 400 pour *eccentricity* soulignant, comme discuté lors de l'évaluation stationnaire, que *eccentricity* a des effets à long terme dus à la non stationnarité.

Nous avons donc opté pour une autre approche basée sur la décroissance exponentielle et nous avons estimé l'α_v moyen de chaque variable (hors variables logiques) v à partir de l'auto-corrélation acf_v pris à $t-15$ pour chacune des 36 trajectoires.

$$\overline{\alpha_v} = \frac{\sum_{i=1}^{36} acf_v(t-15)^{1/15}}{36}$$

A partir de ces α moyens par variable, nous avons déterminé le temps de relaxation par variable τ_v que nous avons moyenné pour obtenir le temps de relaxation $\overline{\tau}$ du jeu de données.

$$\tau_v = \frac{1+\alpha_v}{1-\alpha_v}$$
$$\overline{\tau} = \frac{\sum_{i=1}^{V} \tau_v}{V}$$

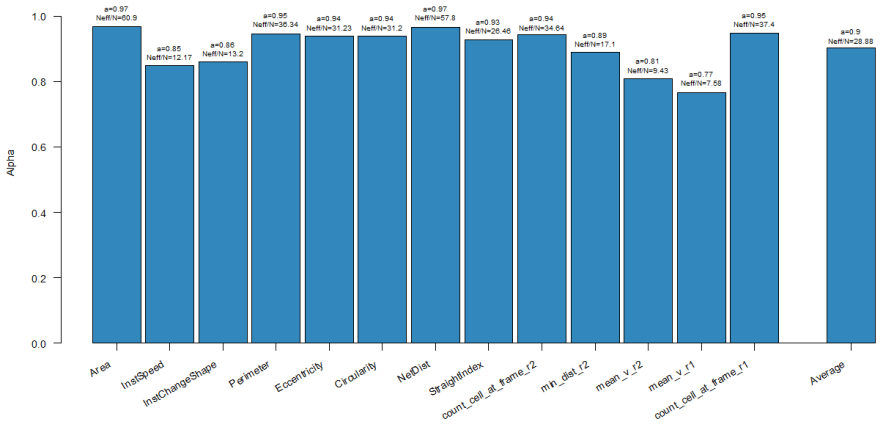


FIGURE 5.27 – **Détermination du nombre de pas temps pour la découverte causale.** Pour chacune des variables, une estimation du α et du temps de relaxation est réalisée.

Nous avons ensuite retenu comme nombre de pas de temps pour la découverte causale τ deux fois le temps de relaxation moyen $\overline{\tau}$ pour s'assurer de couvrir l'ensemble de la dynamique du système.

Pour notre application, le τ obtenu étant de 57, le graphe à reconstruire aurait été composé de $58 \times 15 + 2 = 872$ nœuds, ce qui reste difficilement calculable avec les moyens actuels. Nous avons utilisé le paramètre δ_t pour ne conserver qu'un pas de temps tous les 5 pas de temps, ce qui nous permet de reconstruire un graphe couvrant la dynamique $t-55 \rightarrow t$ avec 11 niveaux d'historique, les niveaux étant espacés de 5 pas de temps.

5.4.4 Résultats

En mode stationnaire, le graphe retourné par tMIIC est un graphe déplié dans le temps sans répétition des arêtes stationnaires, Fig. 5.28a. Ce graphe peut être complété par stationnarité pour visualiser la dynamique complète, Fig. 5.28b.

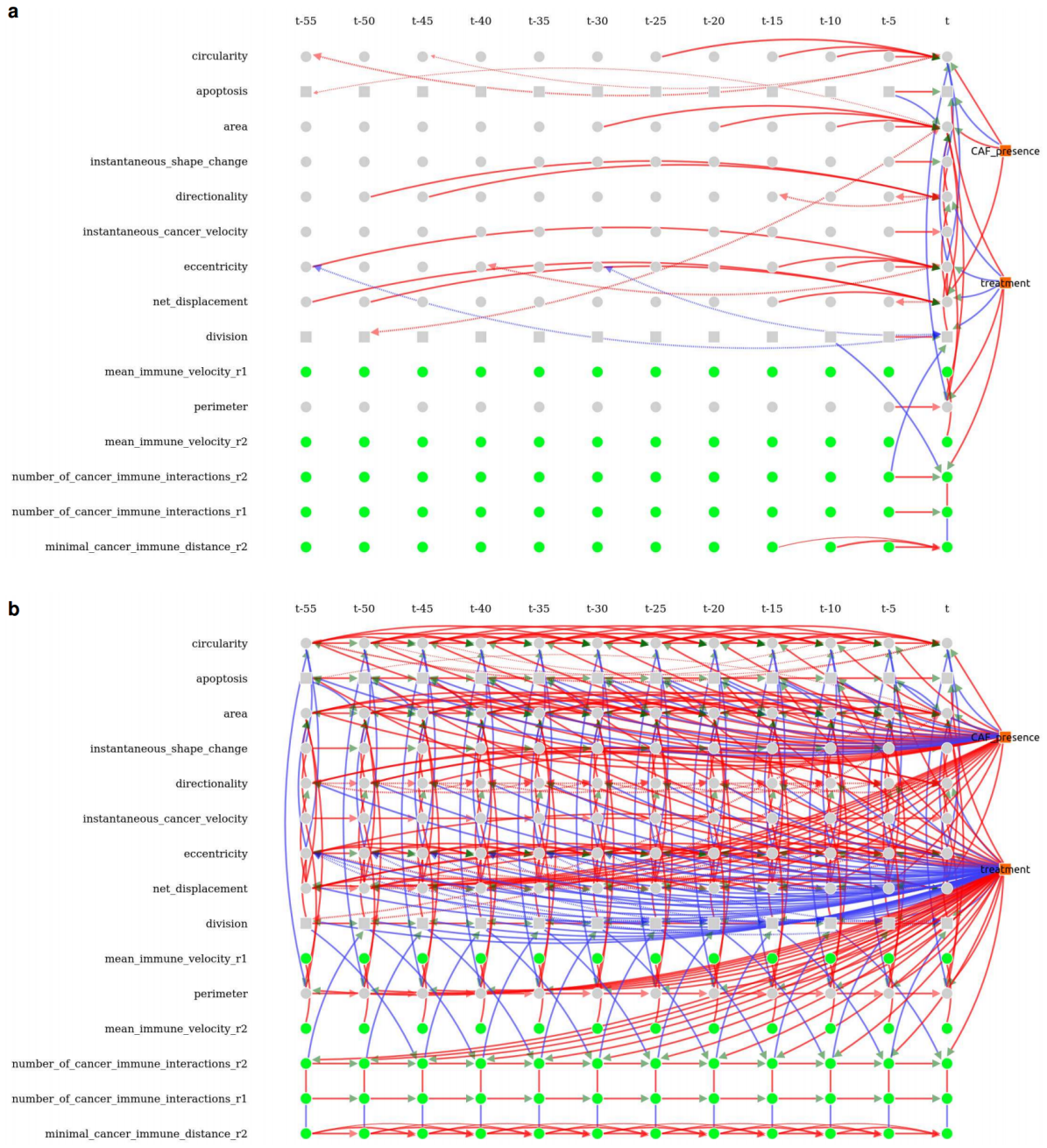


FIGURE 5.28 – **Résultat de la découverte causale par tMIIC.** a Graphe sans répétition des arêtes stationnaires. b Graphe déplié complet.

Pour mieux visualiser le graphe obtenu, une représentation compacte a été implémentée, Fig. 5.29.

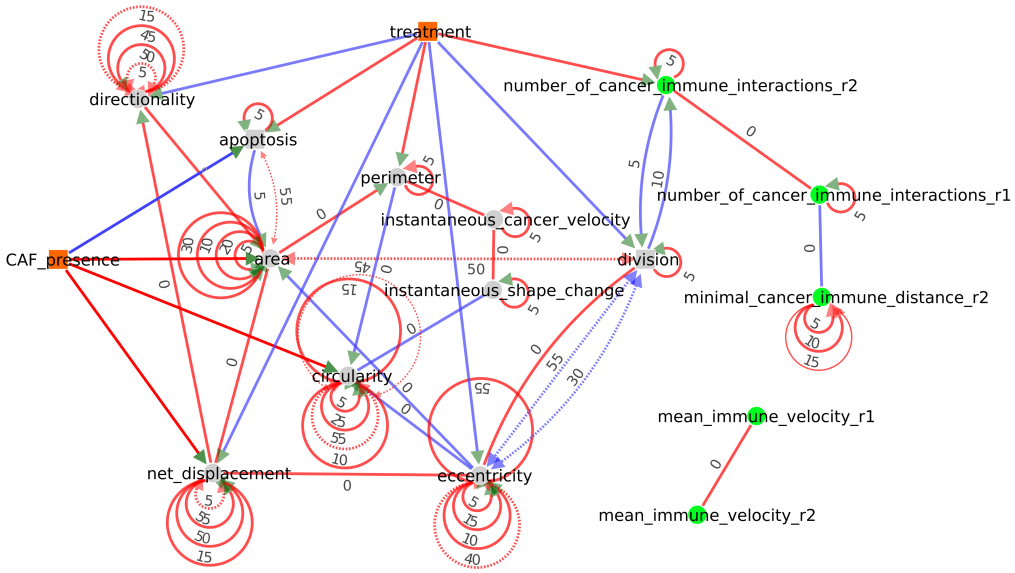


FIGURE 5.29 – **Graphe compact du résultat de la découverte causale par tMIIC.**

Le réseau inféré par tMIIC révèle de nouvelles découvertes biologiquement pertinentes en plus de confirmer les résultats connus d'études antérieures. En particulier, tMIIC découvre que les CAFs inhibent directement l'apoptose des cellules cancéreuses, indépendamment du traitement anticancéreux, Fig. 5.30, alors que des études antérieures ont rapporté que les CAFs réduisaient simplement l'effet du traitement [55]. Sur des éléments connus, tMIIC retrouve que les CAFs stimulent la migration des cellules cancéreuses et augmentent leur aire et circularité.

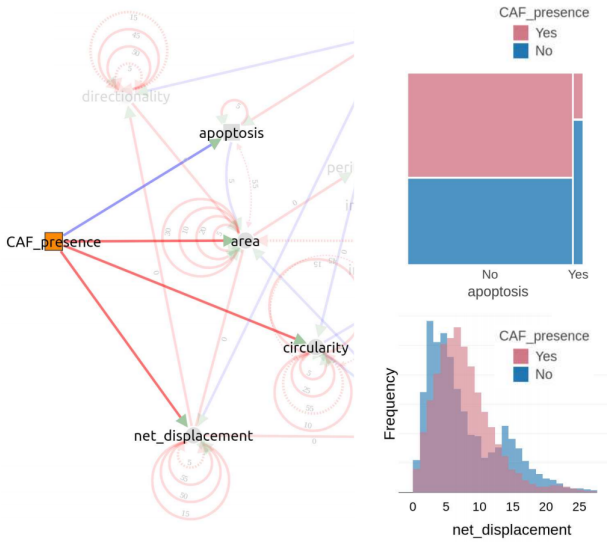


FIGURE 5.30 – **Inhibition directe de l'apoptose des cellules cancéreuses par les CAF.**

tMIIC découvre également que le traitement réduit le périmètre des cellules cancéreuses et inhibe leur migration, Fig. 5.31, ce qui n'avait pas été signalé jusqu'à présent non plus. De plus, tMIIC confirme des résultats connus d'études antérieures, il retrouve notamment que le traitement augmente l'apoptose des cellules cancéreuses et le nombre d'interactions entre cellules cancéreuses et immunitaires, ainsi qu'il diminue le taux de division des cellules cancéreuses.

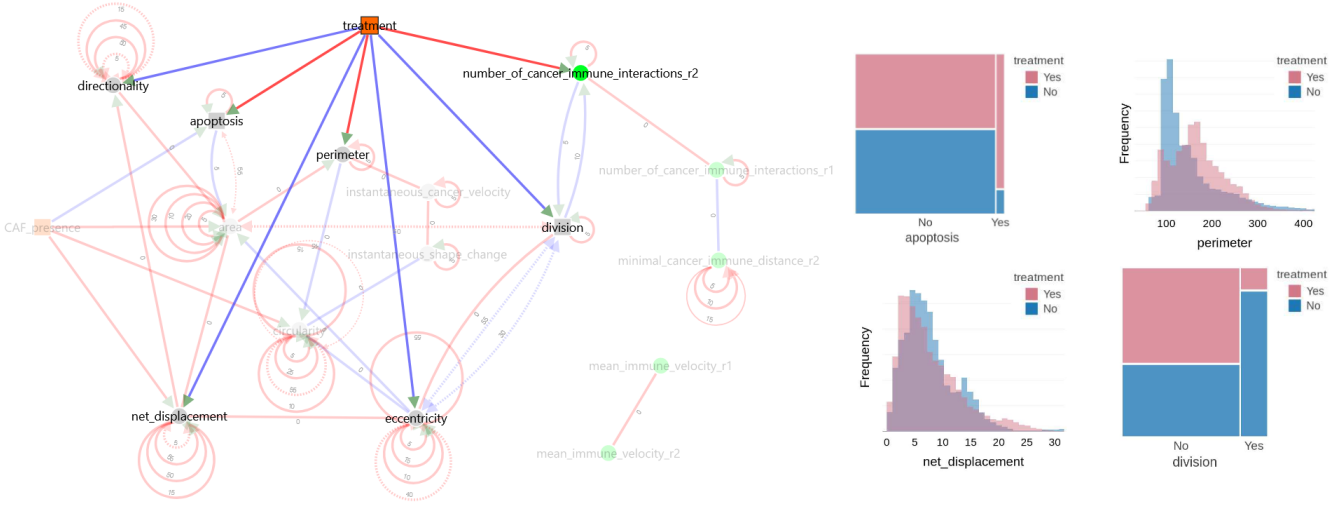


FIGURE 5.31 – **Influence du traitement** sur l'apoptose, la division, le périmètre des cellules cancéreuses et leur migration ainsi que le nombre d'interactions cellules cancéreuses-immunitaires.

Fait intéressant, tMIIC identifie également des effets multiples et éventuellement antagonistes avec des délais différents. Par exemple, tMIIC récupère plusieurs relations antagonistes entre des caractéristiques morpho-dynamiques telles que la division cellulaire et l'excentricité ou l'apoptose et la surface des cellules, Fig. 5.32.

Effectivement, les phases tardives de la division cellulaire sont associées à une augmentation marquée de l'excentricité (arête rouge) mais précédées d'une nette diminution de l'excentricité, une à deux heures avant la cytocinèse (arêtes bleues), une fois que la décision de division a été prise (c'est-à-dire la cause latente probable) et que la cellule duplique en fait son matériel biologique (prophase).

De même, le changement de surface lors de l'apoptose prédit d'abord une diminution peu après l'apoptose (arête bleue) avant une augmentation éventuelle lors de la lyse cellulaire (arête rouge). Enfin, tMIIC découvre que la division cellulaire réduit le nombre d'interactions entre cellules cancéreuses et immunitaires, comme attendu, en les répartissant entre les cellules filles, mais inversement tMIIC trouve que les interactions entre les cellules cancéreuses et immunitaires inhibent la division cellulaire.

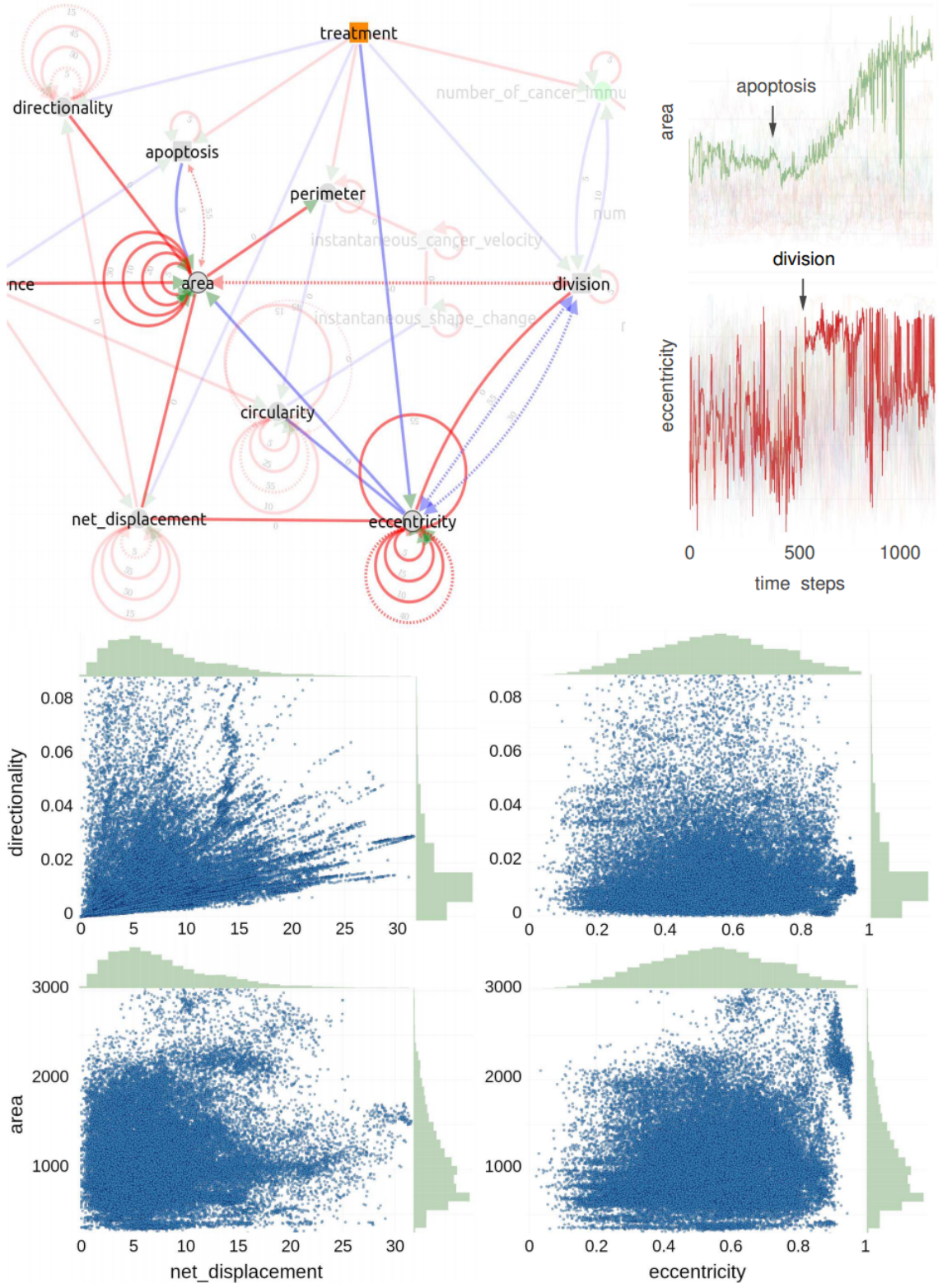


FIGURE 5.32 – **Effets multiples et éventuellement antagonistes avec des délais différents**

6 CausalXtract

Le laboratoire ayant trouvé très intéressants les résultats obtenus en appliquant tMIIC stationnaire aux séquences d'images provenant de cultures cellulaires au microscope, il a été décidé de mettre à disposition de la communauté scientifique une méthode complète permettant, à partir de séquences d'images, d'extraire les variables d'intérêt et reconstruire le graphe temporel causal sous-jacent. Cette méthode, dénommée CausalXtract correspond au pipeline suivant :

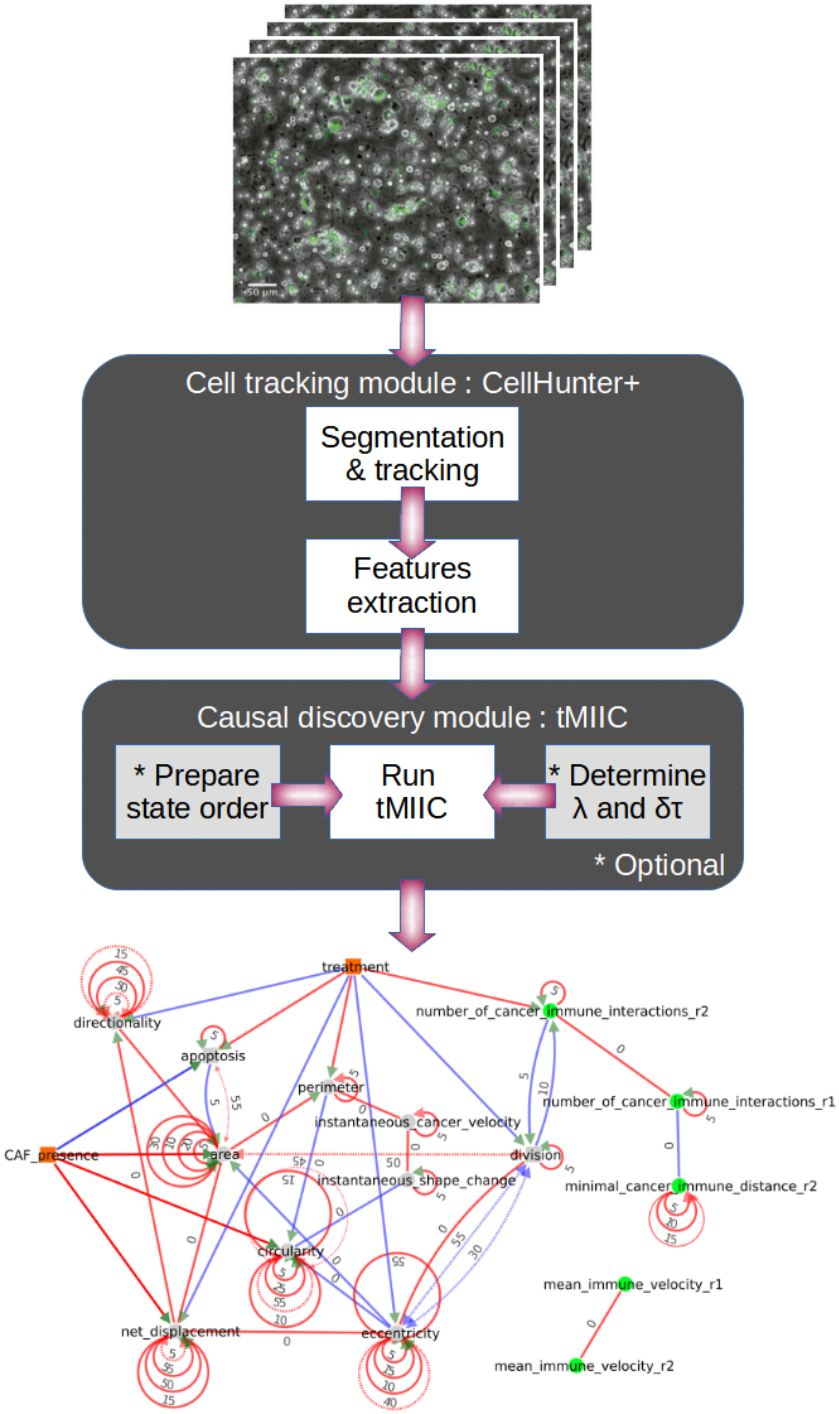


Figure 6.1 – **Pipeline de CausalXtract.**

6.1 Extraction des caractéristiques cellulaires

En utilisant l'expérience acquise lors de l'application de tMIIC sur les données d'expériences sur micro-environnement tumoral reconstitué [55], CausalXtract intègre un module d'extraction des caractéristiques cellulaires CellHunter+ profondemment remanié pour produire des données directement utilisables par tMIIC.

En effet, le module CellHunter+ était initialement construit pour analyser spécifiquement l'expérience sur puce de micro-environnements tumoraux comprenant des cellules immunitaires et des fibroblastes. Il a donc été adapté pour pouvoir être appliqué à des expériences dont l'organisation est différente et il n'est donc plus nécessaire, par rapport à la version initiale, d'avoir des conditions expérimentales (traitement, fibroblastes) et la détection d'interactions avec d'autres cellules a été rendue optionnelle.

CellHunter+ est une méthode qui segmente les cellules, les suit et en extrait les caractéristiques. La segmentation est basée sur la méthode CHT (Circular Hough Transform), le suivi est basé sur l'algorithme de Munkres [56]. Il étend les fonctionnalités de CellHunter [55], qui implémentait la segmentation et le suivi, en ajoutant le module d'extraction de caractéristiques.

Le module "segmentation et tracking" permet d'analyser chaque ROI (Région d'intérêt). Les cellules d'étude principale (dans notre exemple d'application, les cellules cancéreuses) sont placées au centre de la culture et il est possible d'observer les autres cellules (comme par exemple des fibroblastes, des cellules immunitaires).

Les vidéos de ces ROI sont sauvegardées sous forme de matrices où la troisième dimension représente le temps (i.e. le numéro d'image). Les sorties sont les trajectoires des cellules principales d'étude ainsi que, si demandé, les trajectoires d'autres cellules interagissant avec les cellules principales. Une étape supplémentaire permet de corriger le "scintillement" de la trajectoire des cellules principales d'étude lors de leur éventuelle division. Enfin, les trajectoires des cellules principales et secondaires sont utilisées pour calculer les caractéristiques d'intérêt pour chaque ROI.

Bien sûr, avec l'expérience acquise lors de l'application de tMIIC à l'expérimentation initiale, les variables problématiques ont été retirées des sorties produites. CellHunter+ ne fournit donc plus en sortie les variables en contradiction avec la temporalité ou cumulatives (cf 5.4.1) et la sortie de CellHunter+ peut être directement utilisée comme entrée de la découverte causale temporelle.

6.2 Découverte causale temporelle

La découverte causale temporelle a elle aussi été améliorée par rapport à sa forme initiale de tMIIC décrite dans la section 5.1 en ajoutant, notamment, une estimation automatique de la dynamique temporelle du système.

Le principe utilisé pour déterminer automatiquement la dynamique est basé sur la décroissance de la corrélation. Pour cela, nous déterminons d'abord L, la longueur minimale de toutes les trajectoires. L est utilisé comme pas de temps maximum pour calculer l'auto-corrélation $acf_{t,v}$ pour chaque trajectoire t et variable v. Ensuite, nous recherchons pour chaque trajectoire et variable lorsque l'auto-corrélation disparaît et

conservons la moitié du temps de disparition pour récupérer l'auto-corrélation et calculer l'α correspondant.

$$l_{t,v} = \lfloor \frac{min\ (l \mid acf_{t,v}[l] < 0.05)}{2} \rceil$$
$$\alpha_{t,v} = acf_{t,v}[l_{t,v}]^{\frac{1}{l_{t,v}}}$$

Les $\alpha_{t,v}$ sont moyennés sur les trajectoires T et ces moyennes sont utilisées pour déterminer le temps de relaxation τ_v pour chaque variable.

$$\overline{\alpha_v} = \frac{\sum_{t=1}^{T} \alpha_{t,v}}{T}$$
$$\tau_v = \lfloor \frac{1 + \overline{\alpha_v}}{1 - \overline{\alpha_v}} \rceil$$

Pour essayer de couvrir au mieux la dynamique globale du système, le τ retenu pour l'ensemble du jeu de données est le minimum entre la longueur minimale de toutes les trajectoires, le temps de relaxation maximum et deux fois le temps de relaxation moyen :

$$\overline{\tau} = \lfloor \frac{\sum_{v=1}^{V} \tau_v}{V} \rceil$$
$$\tau = min(L, max(\tau_v), 2 * \overline{\tau})$$

Une fois la dynamique à couvrir τ estimée, nous pouvons ensuite déterminer le nombre de couches λ et le nombre de pas de temps δ_t entre chaque couche qui seront utilisés lors de la découverte causale. Pour cela, nous avons introduit un paramètre supplémentaire : le nombre de nœuds maximum dans le graphe laggé final N_{max} pour ajuster, si nécessaire, λ et δ_t tout en couvrant la dynamique τ. En notant N_{Gt} le nombre de nœuds du graphe final, V_c le nombre de variables contextuelles et V_l le nombre de variables laggées :

$$N_{Gt} = (\tau + 1) * V_l + V_c$$

Lorsque $N_{Gt} > N_{max}$, nous réduisons le nombre de couches tout en augmentant le nombre de pas entre chaque couche pour toujours couvrir la dynamique cible :

$$\lambda = max(2, \lfloor \frac{N_{max} - V_c}{V_l} \rfloor)$$
$$\delta_t = max(1, \lfloor \frac{\tau}{\lambda - 1} \rfloor)$$

A ce stade, nous avons presque terminé puisque nous avons déterminé les λ et δ_t qui couvrent la dynamique du système. Cependant, nous nous assurons également qu'en utilisant ces paramètres, nous obtiendrons suffisamment d'échantillons, après décalage dans le temps des données d'entrée, pour effectuer la découverte causale. Pour cela, nous avons posé 1 000 comme nombre minimum d'échantillons souhaité (et si le jeu de données comprend moins de 1 000 pas de temps, 10 % du nombre de pas de temps total) que nous comparons au nombre effectif d'échantillons obtenu après décalage dans le temps des données d'entrée. Si le nombre d'échantillons obtenu est inférieur au nombre cible, δ_t et éventuellement λ sont réduits, avec un message

avertissant l'utilisateur que les paramètres estimés ne couvrent pas la dynamique du système.

Dans la mesure où l'estimation de la dynamique du jeu de données a été modifiée par rapport à celle de la version initiale de tMIIC, l'ensemble des figures initialement prévu a été mis à jour pour la publication de CausalXtract.

La version de tMIIC a, en outre, bénéficié des améliorations apportées lors du développement de CausalXtract puisque tous les changements réalisés ont été reportés dans tMIIC, qui intègre désormais par défaut une estimation automatique de la dynamique temporelle du jeu de données.

6.3 Mise à disposition de la communauté scientifique

Pour conclure sur CausalXtract, l'ensemble du pipeline, tant du côté de l'extraction de caractéristiques cellulaires que de la découverte causale temporelle, a été organisé pour que les scientifiques puissent l'utiliser de la manière la plus facile possible : une fois que les caractéristiques des cellules, qui sont propres à chaque expérimentation, ont été définis, le reste du pipeline requiert un minimum d'intervention humaine avec un ensemble de paramètres qui s'ajustent automatiquement.

Enfin, CausalXtract est accessible à tous sur un dépôt Github public : `https://github.com/miicTeam/CausalXtract`. Un Readme sur le Github reprend les principales fonctions de CausalXtract et une démonstration, reprenant l'ensemble de la publication de CausalXtract, a également été réalisée sous la forme d'un notebook R pour permettre aux scientifiques de se familiariser avec l'outil CaulsalXtract.

7 L'a priori de conséquence

7.1 Principe

Pour préparer la version non stationnaire temporelle de MIIC, nous nous sommes d'abord intéressés à une étape intermédiaire : l'ajout à la version classique de MIIC de l'a priori que certaines variables ne peuvent être que des conséquences, pas des causes, comme les gènes cibles qui ne peuvent pas réguler l'expression d'autres gènes dans les cellules. Du point de vue de la découverte causale, une variable conséquence (*i.e.* une variable qui n'est pas une cause) ne peut jamais être à l'origine ou contribuer à un effet observé sur d'autres variables.

Outre le fait d'être exclue des contributeurs possibles, une variable conséquence doit donc respecter les contraintes suivantes :

- Aucune arête n'est possible entre deux conséquences, puisqu'aucune conséquence ne peut être une cause.
- Les arêtes entre une variable non conséquence et une variable conséquence ont toujours une pointe de flèche vers la conséquence. Par contre, l'extrémité opposée de l'arête dépend du fait d'autoriser ou d'exclure les variables latentes. Si les variables latentes ne sont pas autorisées, nous pouvons être certains que l'extrémité opposée est une cause, en accord avec les hypothèses posées. Si les variables latentes sont autorisées, par contre, il n'est pas possible, a priori, de déterminer l'orientation de l'extrémité opposée qui devra être déduite en utilisant la signature de la causalité.

L'implémentation de la notion de conséquene dans MIIIC a donc plusieurs points communs avec la version temporelle stationnaire : le graphe va pouvoir être élagué en supprimant toutes les arêtes entre conséquences et, d'une manière similaire à la prise en compte de la temporalité, pour la phase d'orientation, il convient de pré-orienter les arêtes avec une pointe de flèche vers les variables conséquences :

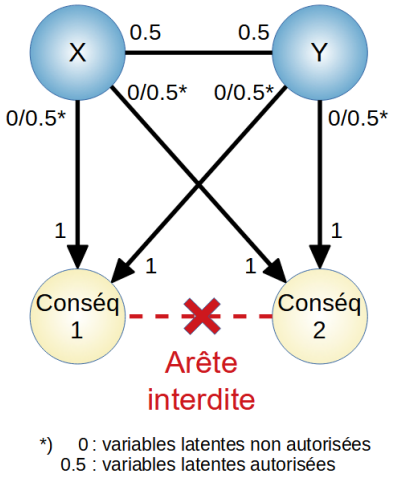


FIGURE 7.1 – **Principe de l'ajout dans MIIC de l'a priori de conséquence.**

7.2 Implémentation dans MIIC

Les modifications apportées correspondent à l'algorithme suivant :

Algorithm 5
Reconstruction causale par MIIC avec consequence (sans variable latente)

Require: $\mathcal{C}$ the set of consequence variables

- Skeleton reconstruction
$\mathcal{G} \leftarrow$ the complete graph on V
for all edges $X - Y \in \mathcal{G}$ **do**
  if ($(X \in C$ and $Y \in C)$ or $\mathrm{I}'(X;Y) \leqslant 0$ **then**
    Delete edge $X - Y$ from $\mathcal{G}$
    Sepset$\{X, Y\} \leftarrow \emptyset$
  else
    Find most contributing node $Z \in \{\mathrm{adj}(X) \cup \mathrm{adj}(Y)\} \setminus C$ which maximizes $R(X, Y; Z|\emptyset)$
  end if
end for
while There is a link $X - Y$ with $R(X, Y; Z|\{U_i\}) > 1/2$ **do**
  for Top link $X - Y$ with highest rank $R(X, Y; Z|\{U_i\})$ **do**
    Expand contributing set $\{U_i\} \leftarrow \{U_i\} + Z$
    if $\mathrm{I}'(X; Y|\{U_i\}) \leqslant 0$ **then**
      Delete edge $X - Y$ from $\mathcal{G}$
      Sepset$\{X, Y\} \leftarrow \{U_i\}$
    else
      Find next most contributing node $Z \in \{\mathrm{adj}(X) \cup \mathrm{adj}(Y)\} \setminus C$ and compute $R(X, Y; Z|\{U_i\})$
    end if
    Sort the rank list $R(X, Y; Z|\{U_i\})$
  end for
end while

- Skeleton orientation
Initialize $X - Y$, $Y \in C$ as $X \rightarrow Y$
Sort list of unshielded triples $\mathcal{L}_c = \{(X, Z, Y)_{X \not\sim Y}\}$ in decreasing order of $|\mathrm{I}'(X; Y; Z|\{U_i\})|$

... orientation from unshielded triples unchanged ...

return $\mathcal{G}$

7.3 Domaine d'application

Le champ d'application que nous avons retenu pour la notion de conséquence est la reconstruction de graphes de régulation de gènes (GRN) basée sur des analyses d'ARN de cellules uniques (scRNAseq). Comparé aux anciennes analyses en "bulk RNA", les analyses en cellule unique fournissent des mesures de l'expression des ARN, donc des gènes transcrits, au niveau de chacune des cellules et ainsi permettent d'identifier des types cellulaires.

FIGURE 7.2 – **Différence entre le séquence d'ARN en bulk et en cellule unique.** Le scRNAseq permet d'obtenir le profil d'expression pour chaque cellule et de déterminer son type.

Le fait de pouvoir disposer d'informations pour chaque type de cellule est primordial pour les applications que nous prévoyons car, d'une part, les expérimentations menées visent généralement à étudier quelques types de cellules parmi toutes celles qui sont séquencées (par exemple, les cellules cancéreuses, les cellules immunitaires) et d'autre part, une reconstruction de réseaux sur l'ensemble des cellules sans les distinguer ne pourrait pas mettre en évidence les réseaux de régulation de gènes spécifiques à chaque type cellulaire. En s'appuyant sur la figure 7.2, cela équivaudrait à inférer un réseau causal sur le smoothie, ce qui ne permettra pas d'identifier les réseaux de régulation réels à l'œuvre dans les framboises et les myrtilles. En conséquence, les reconstructions de graphes causaux se placent en aval de la chaîne de traitement classique des analyses en cellule unique illustré par la figure 7.3, après identification des types cellulaires.

La principale problématique posée pour reconstruire des réseaux causaux sur ce type d'analyse est la dimension du jeu de données puisque les analyses scRNAseq fournissent les niveaux d'expression de chaque gène, soit des dizaines de milliers de gènes, chacun étant une variable pour un nombre d'échantillons allant, lui aussi, jusqu'à plusieurs milliers ou dizaines de milliers de cellules.

En posant l'hypothèse, certes très forte, que tous les gènes sont des conséquences à l'exception des facteurs de transcription, le nombre de gènes qu'il devient possible d'inclure dans le graphe est considérablement augmenté. Alors que, sans utiliser la notion de conséquence, le nombre de gènes qu'il est possible d'analyser se limite à quelques centaines, il devient abordable de reconstruire des graphes avec plusieurs

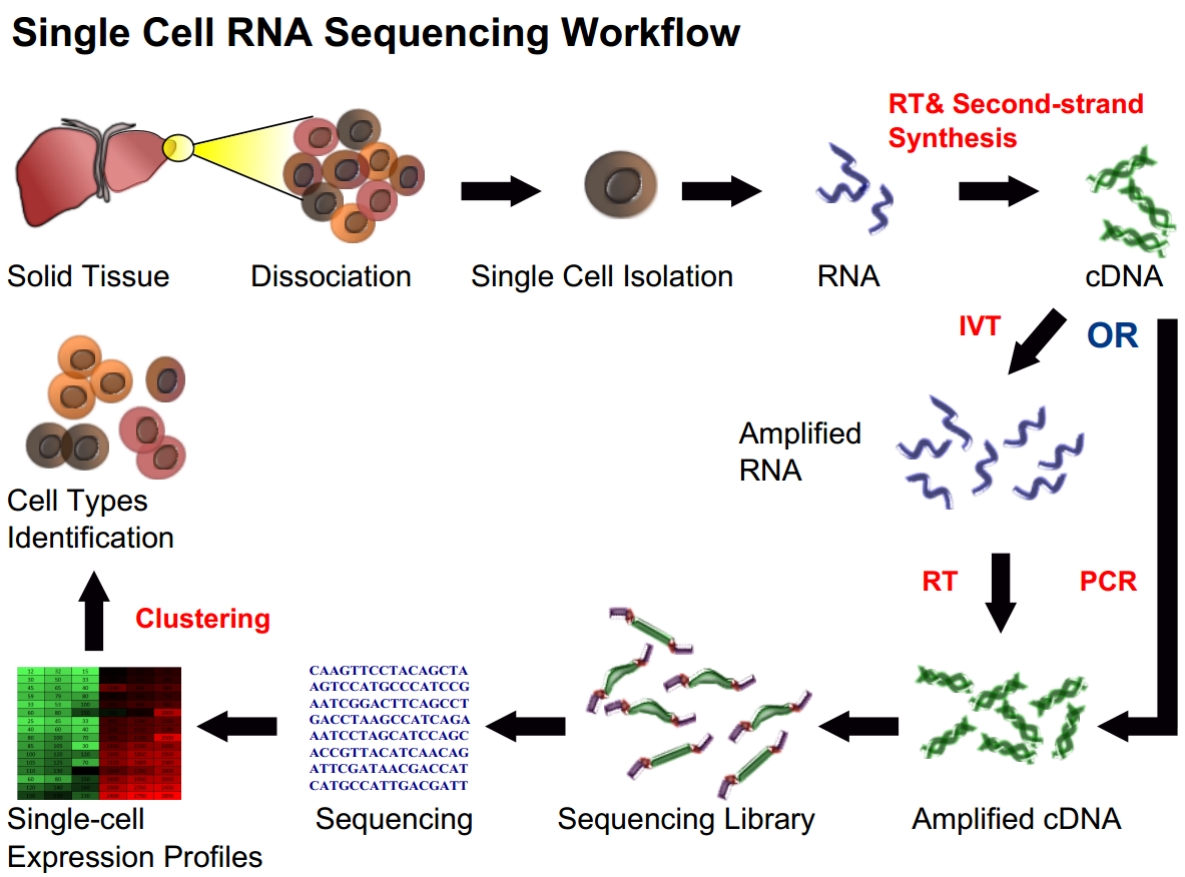


FIGURE 7.3 – **scRNAseq workflow.** Pour notre application, nous nous plaçons en aval du workflow classique d'analyse ARN en cellule unique.

milliers de gènes en posant cette hypothèse que tout gène non facteur de transcription est une conséquence.

7.4 Application

7.4.1 Collaboration avec l'unité Immunité et Cancer

Vu que notre laboratoire fait partie de l'Institut Curie, l'application a très logiquement porté sur une expérimentation dans le domaine de la recherche sur le cancer. Nous avons en effet entamé une collaboration avec Hélène Moreau, chargée de recherche dans l'équipe d'Ana-Maria Lennon-Duménil (unité Immunité et Cancer (U932), Institut Curie). Les résultats de cette collaboration ont été présentés à la conférence "Imaging the Immune System" en juillet 2023 et le poster est disponible en annexes.

Dans l'équipe travaillant avec Hélène Moreau, Zoé Fusilier était en effet intéressée pour comprendre comment, dans leur expérience sur des cellules de cancer du colon, l'administration d'un traitement basé sur une déplétion des macrophages (fig 7.4) pouvait causer un remodelage de la matrice extra-cellulaire qui était associé à l'infiltration de la tumeur par des cellulaires immunitaires (fig 7.5).

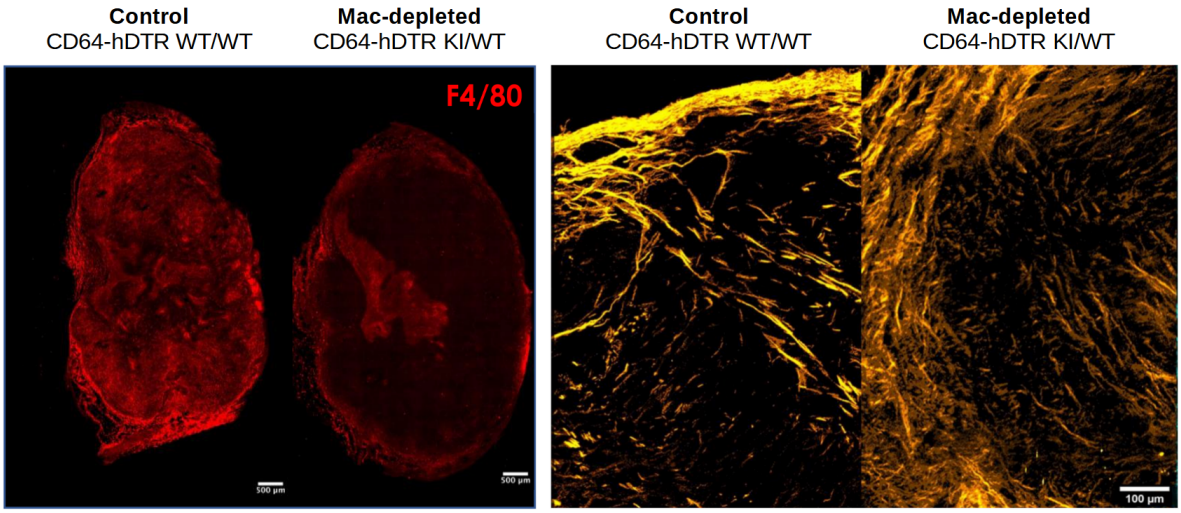


FIGURE 7.4 – **Exemple de déplétion de macrophages.** La déplétion des macrophages (à gauche) entraine une réorganisation de la matrice extra-cellulaire à la périphérie de la tumeur (à droite).

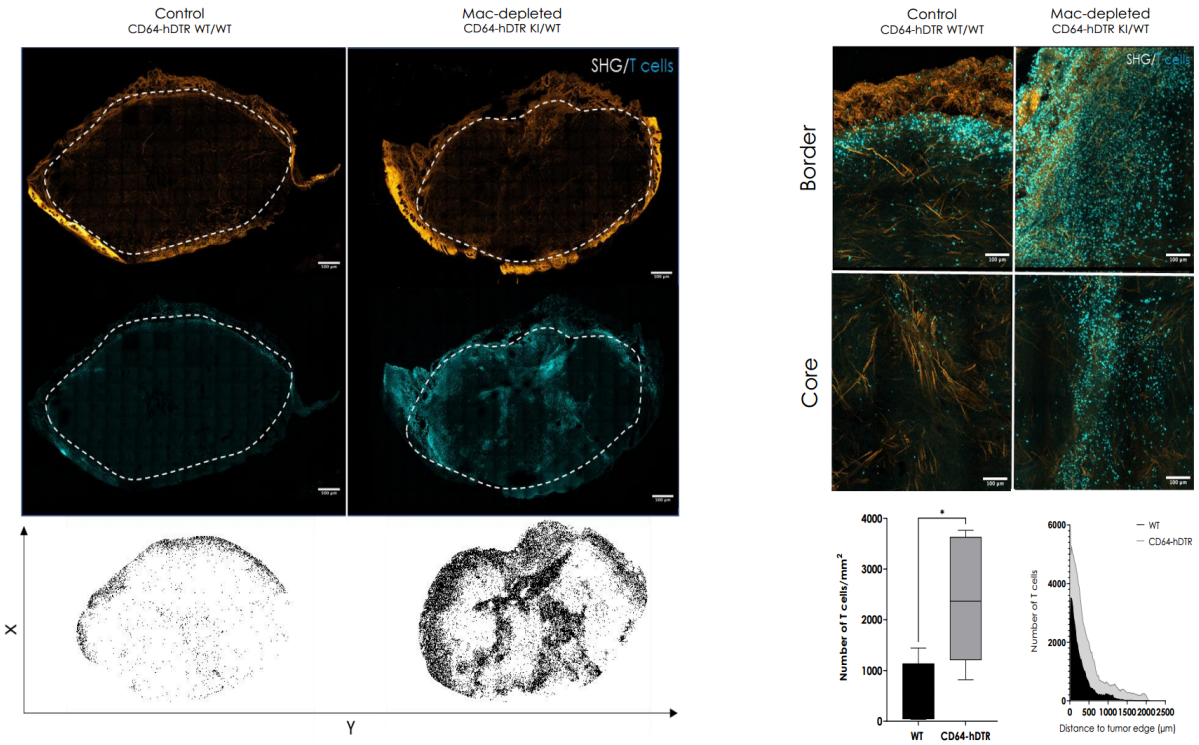


FIGURE 7.5 – **Exemple d'infiltration de cellules immunitaires.** L'infiltration de lymphocytes T au sein de la tumeur corrèle avec le remodelage de la matrice extra-cellulaire.

Si, par des expérimentations, Zoé Fusilier avait déjà identifié certains collagènes (collagène 1 et 3 notamment) comme impliqués dans le remodelage de la matrice extra-cellulaire, la question de la voie par laquelle s'effectuait ce remodelage restait à déterminer et nous avons utilisé MIIC avec la notion de conséquence pour essayer de répondre à cette question.

7.4.2 Données fournies

Les données fournies étaient une analyse en cellules uniques de 14 118 cellules avec 10 865 contrôles et 3 253 cellules provenant d'échantillons traités par déplétion des macrophages. L'étude de ces données avait déjà été réalisée par Zoé Fusilier qui avait identifié 11 types cellulaires :

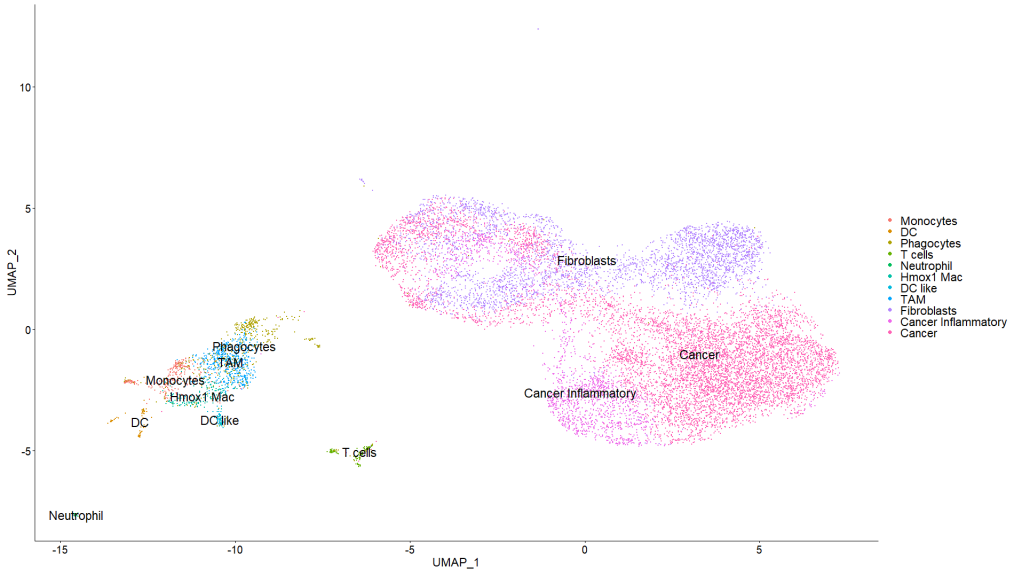


FIGURE 7.6 – **Clusters construits sur l'analyse en cellule unique** avec les 11 types de cellules identifiés.

Sur ces 11 types cellulaires, les deux types les plus importants à étudier étaient les fibroblastes et les cellules cancéreuses qui, fort heureusement, était les cellules les plus représentées dans le jeu de données :

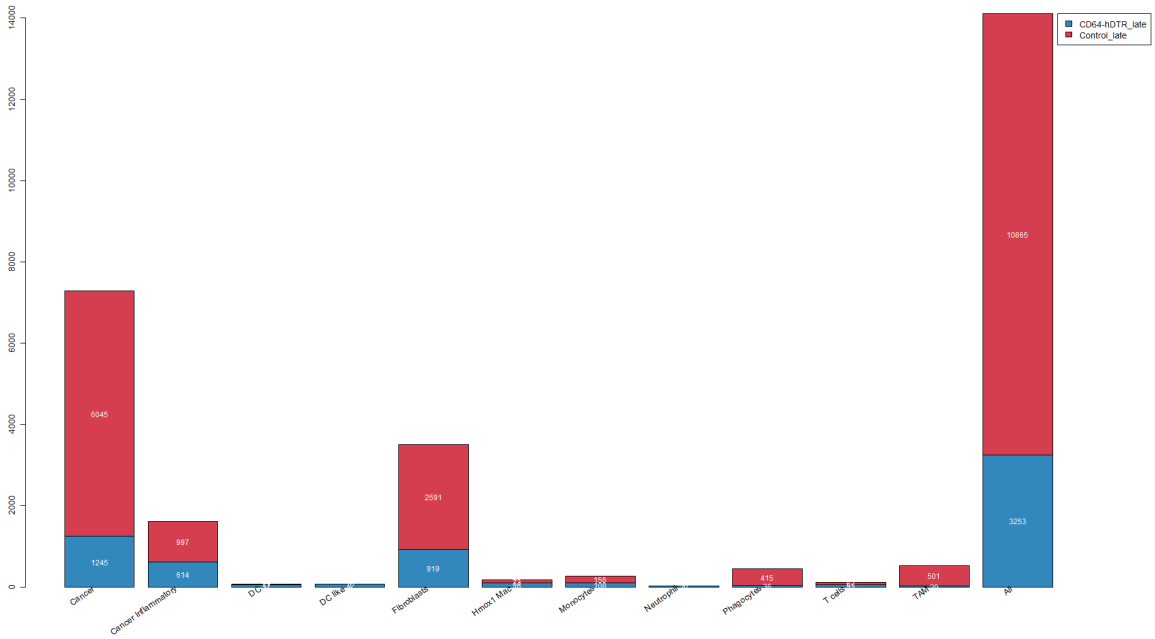


FIGURE 7.7 – **Répartition des cellules par cluster.**

7.4.3 Approche par marqueurs

La première approche mise en place a été classique avec identification de marqueurs (gènes différentiellement exprimés) pour chaque type cellulaire, par rapport à la condition contrôle vs traitement selon le pipeline décrit ci-dessous :

- Utilisation de l'analyse scRNAseq réalisée pour obtenir des clusters et les types cellulaires associés comme présenté dans la figure 7.3
- Pour chaque type cellulaire, identification de gènes marqueurs en fonction de la variable traitement versus contrôle
- Filtrage des marqueurs selon leur p-valeur ajustée et leur variation d'expression moyenne en fonction du nombre des gènes souhaité dans le réseau
- Ajout d'une liste de gènes d'intérêt fournie par les biologistes
- Exclusion des gènes ribosomiaux ou mitochondriaux (qui sont souvent considérés comme "ordure")
- Extraction des données d'expression pour les gènes retenus avec un filtrage sur le taux d'expression >= 1% (pour éviter les valeurs constantes non informatives)
- Marquage des gènes non facteurs de transcription comme conséquence
- Reconstruction du graphe causal par MIIC

L'application de la méthode Seurat *findMarkers* avec ses valeurs par défaut a détecté très peu de marqueurs sur les types cellulaires d'intérêt mais avec d'excellentes p-valeurs :

Type de cellule	Nb cells		p-value filter									
	CD64	Control	1	0.99	0.5	0.2	0.1	0.05	0.01	0.001	1E-06	1E-09
Cancer	1245	6045	11	11	11	11	11	11	11	11	11	11
Fibroblasts	919	2591	18	18	18	18	18	18	18	18	18	18

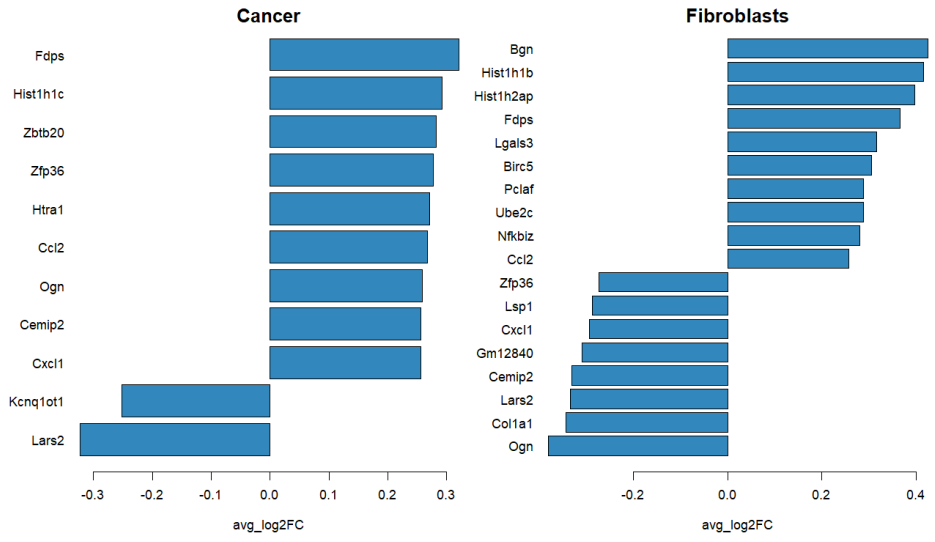


FIGURE 7.8 – Marqueurs pour les types Cancer et Fibroblastes.

Une fois soumis sur le serveur, la version MIIC avec conséquence nous a permis d'obtenir les réseaux suivants :

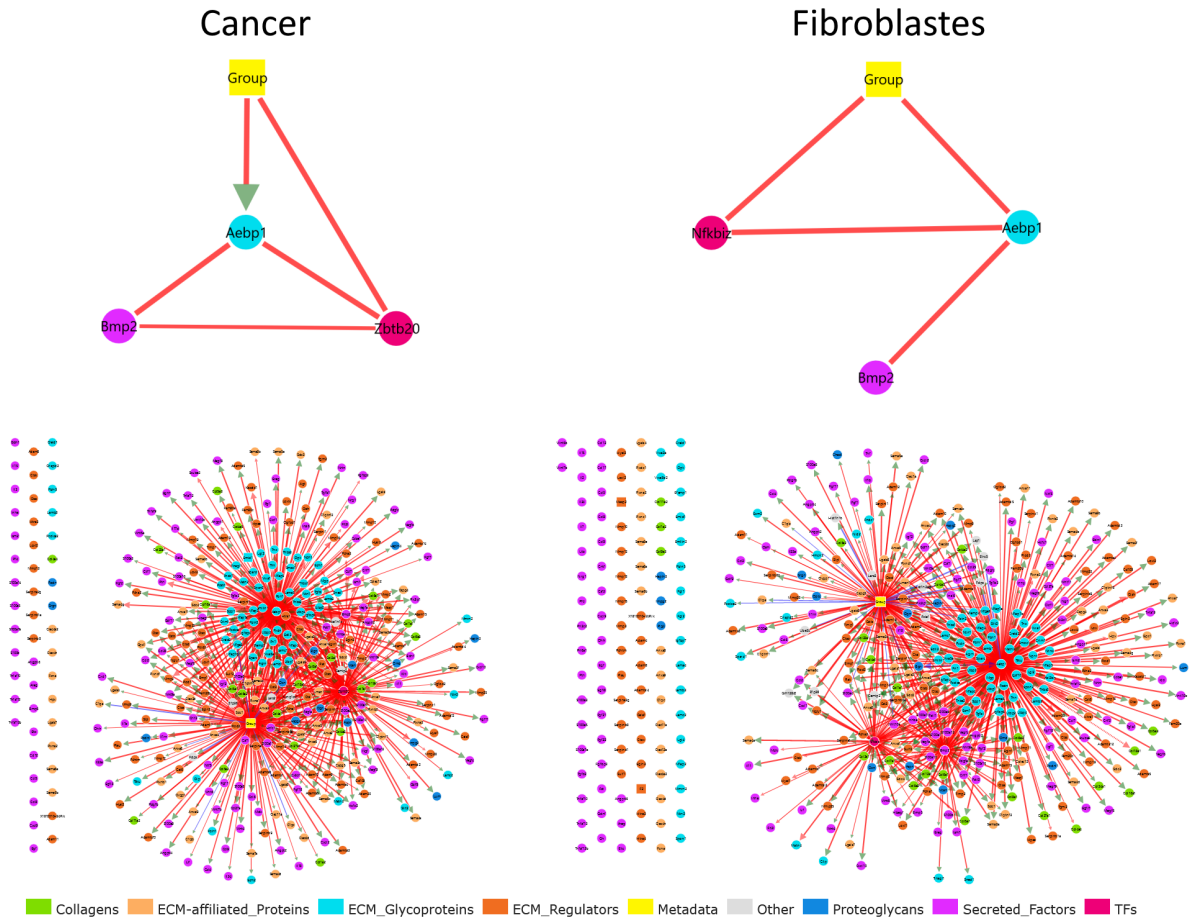


FIGURE 7.9 – **Réseaux inférés par MIIC conséquence** avec en haut, les réseaux avec uniquement les facteurs de transcription. En bas, les réseaux complets.

Cette première approche nous a montré que l'a priori de connaissance, utilisé selon cette approche, était très efficace pour réduire le temps de traitement puisque le graphe du groupe *Cancer* a nécessité moins de 7 minutes pour 381 nœuds et 7 290 échantillons alors qu'un graphe complet (sans utiliser la notion conséquence) de 301 nœuds a pris plus de 22 heures pour fournir le résultat.

Autre point notable, même s'il était attendu, MIIC identifie parmi les gènes d'intérêt des biologistes, un ensemble de gènes sans aucune connexion que ce soit avec le traitement ou les facteurs de transcription, indiquant l'absence d'effet du traitement sur ces gènes. Tous ces gènes sans lien pourraient être retirés des analyses ultérieures pour se concentrer sur les gènes impactés par le traitement.

Ces résultats ont beaucoup intéressé les biologistes puisqu'ils ont permis d'identifier des facteurs de transcription qui étaient des causes possibles des changements de collagènes comme le montre la figure 7.10.

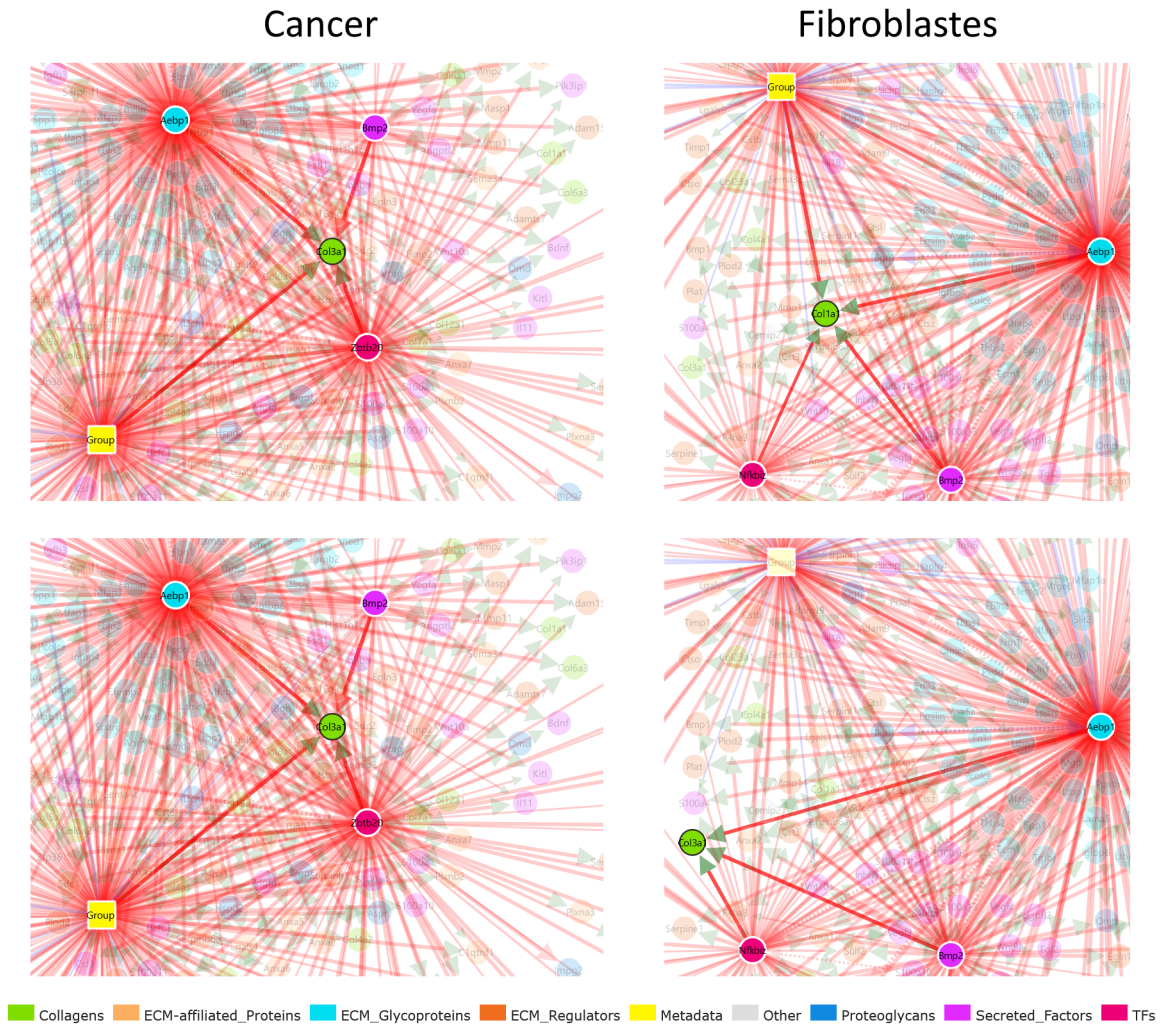


FIGURE 7.10 – **Focus sur les collagènes 1 et 3**

D'autres résultats, en terme de méthodes, sont également intéressants car l'augmentation du nombre de facteurs de transcription permet le déplacement d'arêtes indiquant la présence d'une cause latente vers la cause probable lorsque des facteurs de transcription supplémentaires sont ajoutés comme illustré sur la figure 7.11.

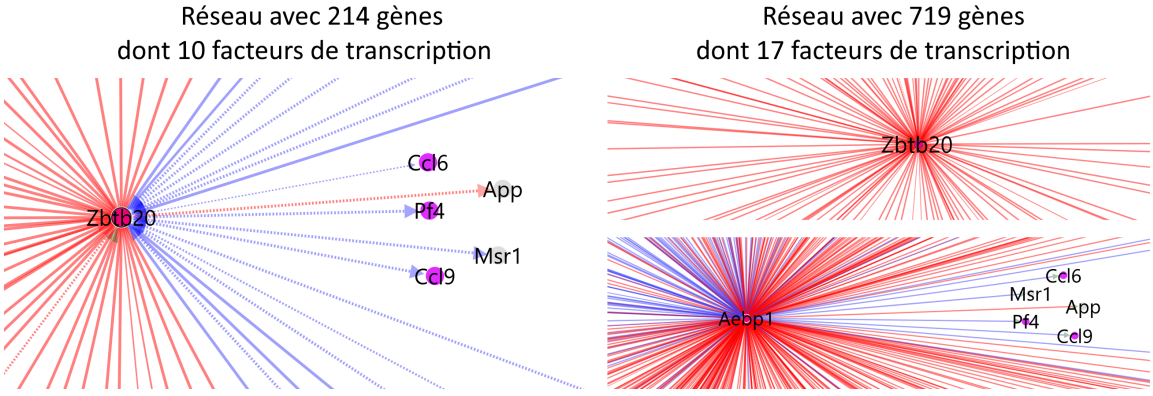


FIGURE 7.11 – **Exemple de transfert d'arête indiquant une cause latente vers la cause probable lors de l'ajout de facteurs de transcription dans le réseau.**

7.4.4 Approche par changement du niveau d'expression

Cependant, même si nous avions des résultats intéressants, cette première approche basée sur les marqueurs était encore loin de notre objectif d'inclure plusieurs milliers de gènes dans nos reconstructions causales. Lors de nos échanges avec les biologistes, il est apparu que les facteurs de transcription n'ont besoin que d'une faible variation d'expression pour agir, par effet de levier, sur les gènes cibles et l'approche par marqueur n'était donc pas la plus effective pour les facteurs de transcription.

Le nombre de facteurs de transcription chez la souris, l'animal modèle pour ces expériences, est trop important pour tous les conserver puisqu'il y a plus de 1 500 facteurs de transcription connus chez la souris. Un filtrage restait donc nécessaire et nous avons opté en seconde approche pour sélectionner les facteurs de transcription selon leur changement de niveau d'expression (log fold change).

Le principe peut être résumé de la façon suivante :

- Utilisation de l'analyse scRNAseq réalisée pour obtenir des clusters et les types cellulaires associés présentés dans la figure 7.3
- Pour chaque type cellulaire, calcul du changement du niveau d'expression (log fold change) des facteurs de transcription en fonction de la variable traitement versus contrôle
- Filtrage des N facteurs de transcription ayant les plus grandes variations d'expression
- Ajout de la liste de gènes d'intérêt fournie par les biologistes
- Ajout des gènes marqueurs
- Exclusion des gènes ribosomiaux ou mitochondriaux (qui sont souvent considérés comme "ordure")
- Extraction des données d'expression pour les gènes retenus avec exclusion des gènes constants (non informatifs)
- Marquage de tous les gènes comme conséquence à l'exception des N facteurs de transcription ayant les plus grandes variations d'expression
- Reconstruction du graphe causal par MIIC

Dans la mesure où nous avons plus de 1 500 facteurs de transcription et plus de 1 000 gènes d'intérêt, la taille des réseaux inférés correspondait bien plus à notre objectif, puis nous avons obtenu 1 997 gènes pour le groupe *Cancer* et 2 066 pour le groupe *Fibroblastes* représentés sur la figure 7.12. Comme précédemment, nous pouvons observer un ensemble de gènes sans aucune connexion que ce soit avec le traitement ou les facteurs de transcription, indiquant l'absence d'effet du traitement sur ces gènes. Tous ces gènes sans lien pourraient être retirés des analyses ultérieures pour se concentrer sur les gènes impactés par le traitement.

A ce stade, les améliorations que j'ai apportées sur le serveur spécifiquement pour la notion de conséquence ont été très utiles : ces développements permettant en effet de visualiser par partie le réseau obtenu et donc d'observer plus facilement les sous-réseaux autour de chaque facteur de transcription comme indiqué sur la figure 7.13.

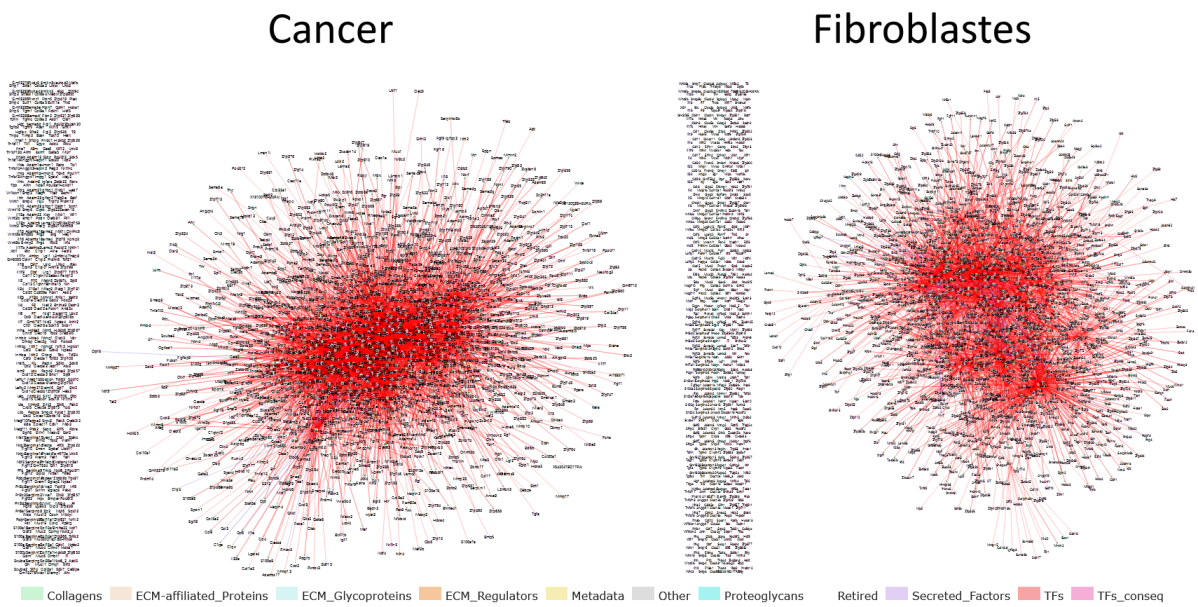


FIGURE 7.12 – **Exemple de réseaux avec 2000 variables.**

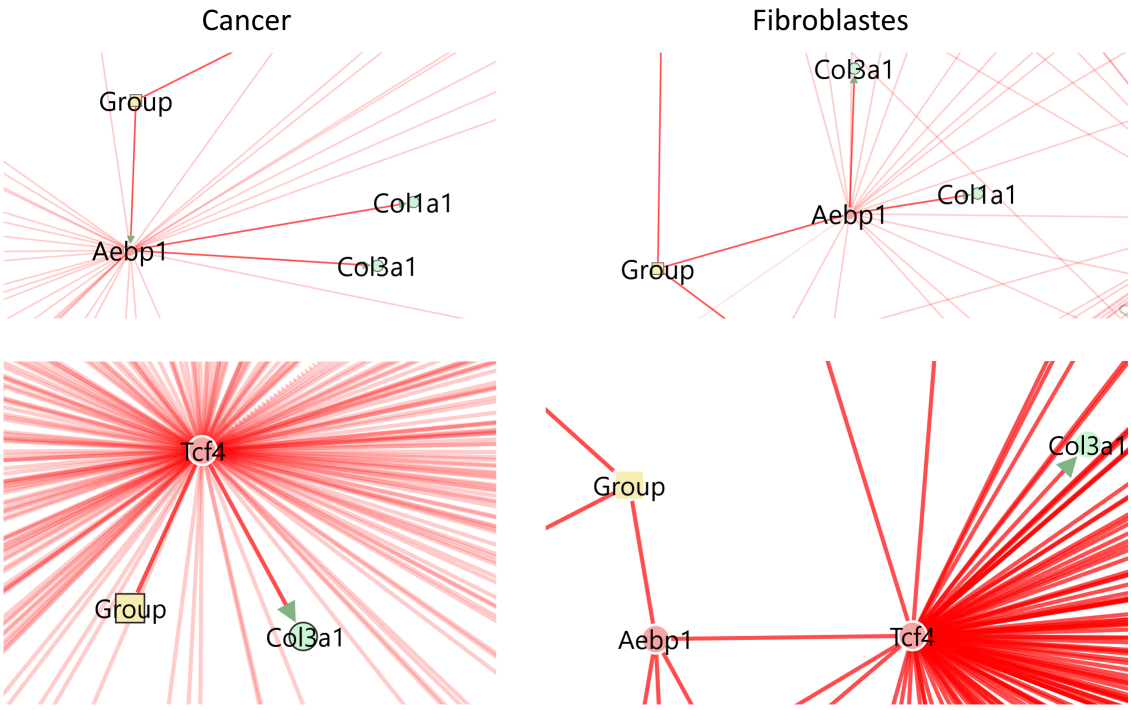


FIGURE 7.13 – **Focus sur des sous-réseaux autour de $Aebp1$ et $Tcf4$ pour visualiser les liens vers les collagènes 1 et 3.**

Si les réseaux obtenus confirment l'importance d'*Aebp*1, qui avait déjà été identifié auparavant, l'ajout de facteurs de transcription au réseau révèle également l'importance du facteur de transcriptions $Tcf4$, notamment dans le cas du cluster *Cancer*. Dans ce réseau, $Tcf4$ revient souvent comme contributeur le plus important pour conditionner les arêtes du traitement avec les autres variables. Par exemple, $Tcf4$ explique > 99 % et 85 % de l'information entre la variable traitement versus contrôle avec, respectivement, les collagènes 1 et 3. De même, $Tcf4$ permet d'expliquer 100 % et 90 % de l'information partagée entre le traitement et, respectivement, *Bmp*2 et *Zbtb*20, alors que ces facteurs de transcription apparaissaient comme liés directement au traitement dans les premiers réseaux n'ayant qu'un petit nombre de marqueurs.

7.4.5 Approche par information mutuelle

Dans la mesure où nous avons pu constater qu'un nombre important de gènes n'était pas connecté dans le réseau obtenu, nous avons envisagé d'autres moyens pour trouver un sous-ensemble de gènes d'intérêt. Plutôt que de rechercher des gènes différentiellement exprimés ou d'utiliser des gènes fournis par les biologistes, nous avons retenu l'information mutuelle comme alternative à la sélection, ce qui est, en outre, plus cohérent avec MIIC qui est basée sur l'information mutuelle.

Pour cela, nous avons donc filtré les gènes constants (qui ne contiennent aucune information) puis calculé l'information mutuelle de chaque gène restant avec la variable traitement versus contrôle. Sur notre application, le filtrage sur les gènes constants a permis d'exclure un tiers des gènes de départ et le calcul de l'information mutuelle a permis d'identifier un autre tiers ne partageant pas d'information avec la variable traitement versus contrôle, ce qui nous laisse une dizaine de milliers de gènes potentiellement intéressants.

Néanmoins, un réseau d'une dizaine de milliers de gènes restant encore actuellement hors de portée, il nous a fallu appliquer une sélection parmi ces gènes possibles. Si, en première approche, nous avons opté pour fixer manuellement un seuil d'information mutuelle pour sélectionner les facteurs de transcription et les gènes, nous avons ensuite opté pour choisir les X facteurs de transcription ayant la plus forte information mutuelle. Les autres gènes ont été sélectionnés parmi ceux ayant au moins le même niveau d'information mutuelle avec la variable traitement versus contrôle que les facteurs de transcription retenus.

L'algorithme a donc été modifié de la manière suivante :

- En entrée, données scRNAseq avec types cellulaires et variable traitement versus contrôle
- Pour chaque type cellulaire, sélection de X (à déterminer selon le matériel) facteurs de transcription en fonction de l'information mutuelle avec la variable traitement versus contrôle.
- Pour chaque type cellulaire, sélection des gènes | $MI_{gene} >= min(MI_{TF\ retenus})$
- Marquage de tous les gènes comme conséquence à l'exception des facteurs de transcription
- Reconstruction du graphe causal par MIIC

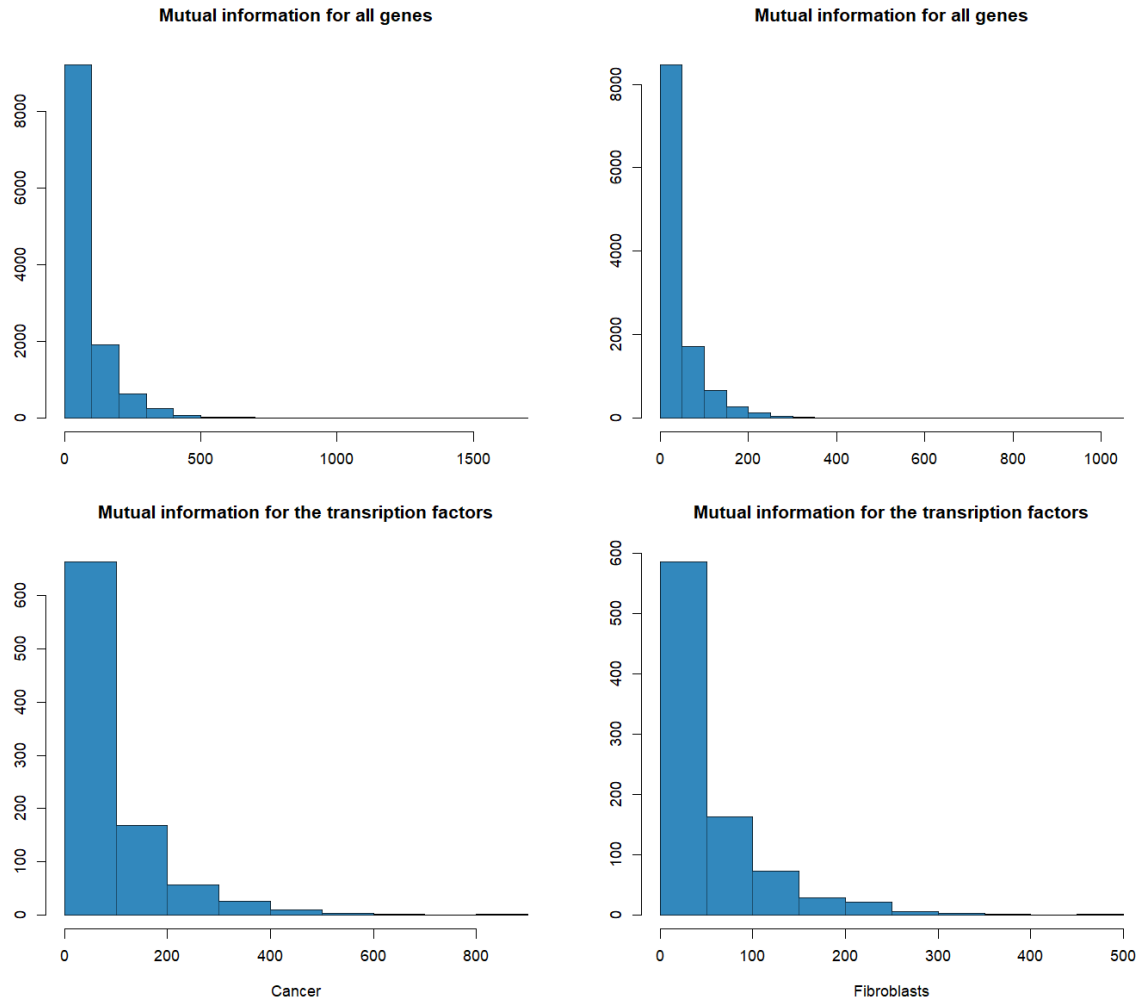


FIGURE 7.14 – **Information mutuelle entre la variable traitement versus contrôle avec les gènes et les facteurs de transcription.**

De manière tout à fait intéressante, les facteurs de transcription retenus diffèrent entre la sélection par le changement du niveau d'expression et selon l'information mutuelle, puisque qu'un quart à un cinquième des facteurs de transcription sélectionnés ne sont pas communs.

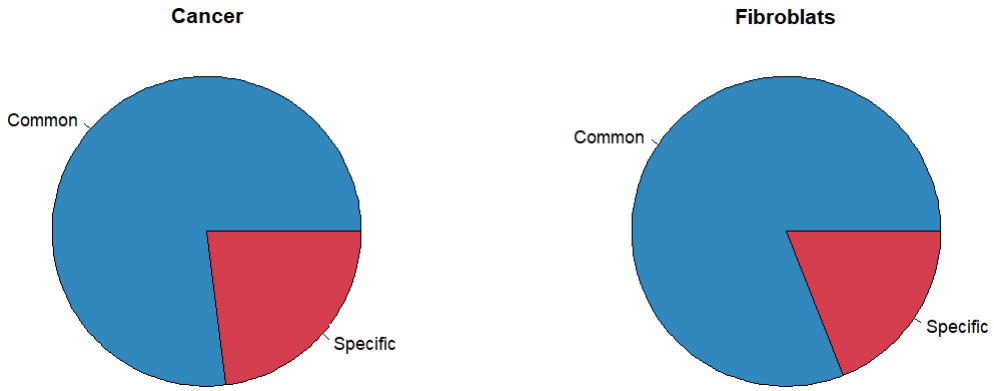


FIGURE 7.15 – **Proportions des facteurs de transcription communs et spécifiques selon la méthode de sélection pour les types cellulaires *Cancer* et *Fibroblastes*.**

A ce stade, il est difficile d'évaluer quelle méthode de sélection est la plus pertinente. Les réseaux obtenus en sélectionnant par changement du niveau d'expression et information mutuelle retrouvent tous deux les facteurs de transcription majeurs *Aebp*1 et *Tcf* 4 et les chemins entre la variable traitement versus contrôle et les collagènes sont très similaires.

En terme de différence entre les deux méthodes, il semblerait que la sélection par l'information mutuelle produise un graphe plus centré autour de la variable expérimentale traitement versus contrôle (22 facteurs de transcription liés à cette variable dans le cluster *Fibroblastes*) par rapport au changement du niveau d'expression (20 facteurs de transcription liés), ce qui est assez logique. A l'inverse, la sélection par changement du niveau d'expression semble permettre d'identifier des facteurs de transcription plus lointains de la variable expérimentale mais qui peuvent être utiles pour analyser les interactions avec des gènes d'intérêt, comme les collagènes dans cet exemple. Sur la figure 7.16, nous pouvons constater que le facteur de transcription *Ybxr*3 explique 5 % de l'information entre *Group* et *Col*1*a*1. De même, les collagènes comportent plus d'arêtes avec des facteurs de transcription par rapport à la sélection utilisant l'information mutuelle.

X	Y	Log fold change						Mutual information					
		Edge Kept	ai	MI	CMI	Complexity	CMI'	Edge Kept	ai	MI	CMI	Complexity	CMI'
Group	Col1a1	O	Aebp1 [83%];Tcf4 [9%];Ybx3 [5%]	289	28	20	8	O	Aebp1 [83%];Tcf4 [9%]	289	48	26	22
	Col3a1	N	Aebp1 [93%];Tcf4 [7%]	340	0	0	0	N	Aebp1 [93%];Tcf4 [7%]	340	0	0	0
	Aebp1	O	NA	512	512	25	488	O	NA	512	512	25	488
	Tcf4	N	Aebp1 [81%];Egr1 [15%]; Gatad2b [2%];Elk3 [1%]; Nr2f1 [<1%];Kdm5b [<1%]	352	0	0	0	N	Aebp1 [81%];Egr1 [15%]; Gatad2b [2%];Elk3 [1%]; Nr2f1 [<1%];Kdm5b [<1%]	352	0	0	0
Col1a1	Aebp1	O	NA	467	467	76	391	O	NA	467	467	76	391
	Zfp36l1	O	Aebp1 [70%]	380	141	48	93	O	Aebp1 [70%]	380	141	48	93
	Lyar	O	NA	80	80	17	63	-	-	-	-	-	-
	Kat6b	O	Tcf4 [85%]	247	71	42	29	-	-	-	-	-	-
Col3a1	Tcf4	O	NA	1088	1088	131	957	O	NA	1088	1088	131	957
	Aebp1	O	NA	945	945	105	840	O	NA	945	945	105	840
	Trps1	O	NA	866	866	128	738	O	NA	866	866	128	738
	Zfp36l1	O	NA	821	821	95	726	O	NA	821	821	95	726
	Six1	O	NA	495	495	83	412	O	NA	495	495	83	412
	Kdm5b	O	NA	473	473	63	409	O	NA	473	473	63	409
	Kat6b	O	NA	421	421	67	354	-	-	-	-	-	-
	Sp140	O	NA	406	406	62	344	O	NA	406	406	62	344
	Rorb	O	NA	405	405	65	339	O	NA	405	405	65	339
	Hbp1	O	NA	370	370	47	323	O	NA	370	370	47	323
	Hoxc8	O	NA	381	381	61	321	-	-	-	-	-	-
	Dnmt3a	-	-	-	-	-	-	O	NA	209	209	38	170
	Zbp1	O	Sp140 [66%]	315	116	28	87	O	Sp140 [66%]	315	116	28	87
	Hic1	O	Group [60%]	258	112	27	85	O	Group [60%]	258	112	27	85
	Mxd3	O	NA	107	107	32	75	-	-	-	-	-	-
	Nr2f1	O	Tcf4 [91%]	537	82	39	43	O	Tcf4 [91%]	537	82	39	43
	Lyar	O	Mxd3 [34%]	69	46	13	33	-	-	-	-	-	-

FIGURE 7.16 – **Comparaison des résultats obtenus avec la sélection par log fold change et par l'information mutuelle sur le cluster *Fibroblastes*.**

7.5 Mise à disposition de la communauté scientifique

Bien que la reconstruction de réseaux de régulation de gènes soit prometteuse et ait montré son utilité en identifiant des gènes importants dans le cas de notre application, elle devra encore mûrir pour pouvoir être mise à disposition de la communauté scientifique comme cela sera exposé dans le chapitre suivant.

A l'inverse, l'a priori de conséquence, qui est pleinement fonctionnel, a été intégré dans le package R mis à disposition de la communauté et est intégré dans le github public : `https://github.com/miicTeam/miic_R_package`.

MIIC avec conséquence est également mis à disposition via le serveur accessible à tous depuis internet à l'adresse `https://miic.curie.fr/workbench.php`.

8 Travaux futurs

8.1 La version temporelle non stationnaire

Même si mon objectif premier au sein de l'équipe du Dr. Hervé Isambert était l'implémentation d'une version temporelle stationnaire de MIIC, j'aurai beaucoup aimé pouvoir l'étendre au cas non stationnaire. Cependant, malgré que le principe et les pré-requis soient établis, la version temporelle non stationnaire reste encore à l'état de travail futur.

Le principe envisagé pour cette version non stationnaire est d'abord d'aligner les trajectoires sur un point commun. Ce point commun peut être le début ou la fin des trajectoires ou encore un évènement sur une variable d'intérêt comme, par exemple, le moment d'injection d'une drogue, un seuil de température atteint, ... Toutes les trajectoires devraient être harmonisées en nombre de pas de temps et l'objectif serait de reconstruire un graphe déplié dans le temps non stationnaire.

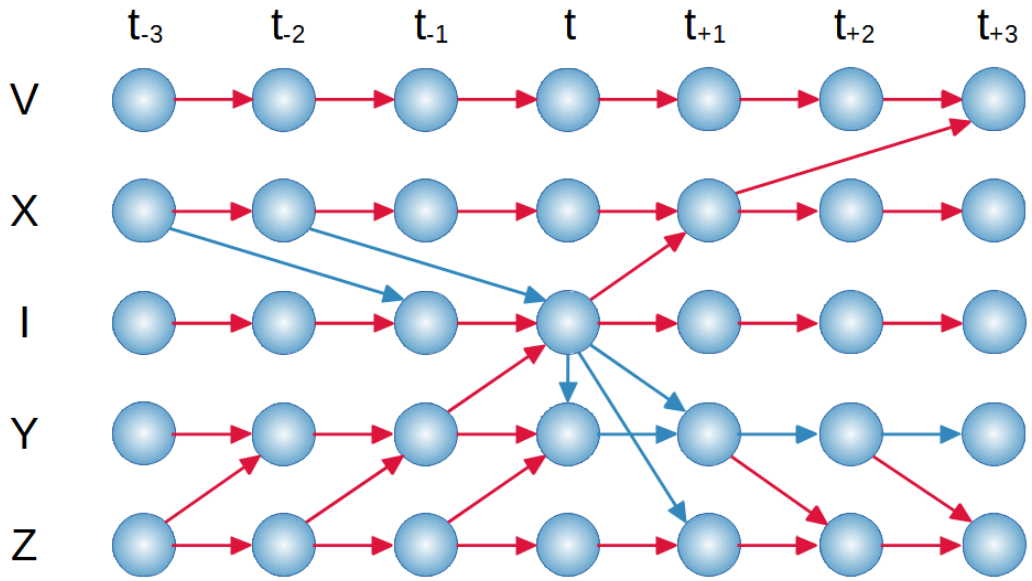


Figure 8.1 – **Exemple fictif de graphe attendu.** Le graphe déplié dans le temps non stationnaire permettrait d'identifier des effets ponctuels, voire opposés à différents pas de temps, ici avec un graphe centré sur une variables d'intérêt I.

Comme nous l'avons évoqué dans les problématiques, pour réaliser une version non stationnaire, il faut gérer la notion d'exclusion des contributeurs futurs. Pour cela, nous prévoyons de ré-utiliser la notion de conséquence introduite dans le chapitre précédent en la rendant variable selon le temps.

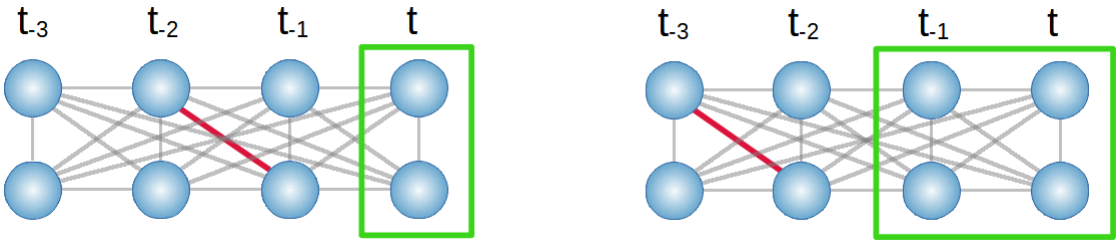


FIGURE 8.2 – **Exclusion de contributeurs en fonction du temps.** Selon l'arête évaluée, les contributeurs à exclure sont différents : avec une arête $t_{-2} - t_{-1}$, les nœuds à t peuvent être considérés comme étant uniquement conséquences du niveau t_{-1} et donc exclus des contributeurs possibles. Pour une arête $t_{-3} - t_{-2}$, l'ensemble des nœuds à t et t_{-1} ne peuvent qu'être des conséquences du niveau t_{-2}.

La phase d'orientation sera similaire à la phase d'orientation de la version stationnaire sans, bien sûr, de duplication des arêtes par stationnarité.

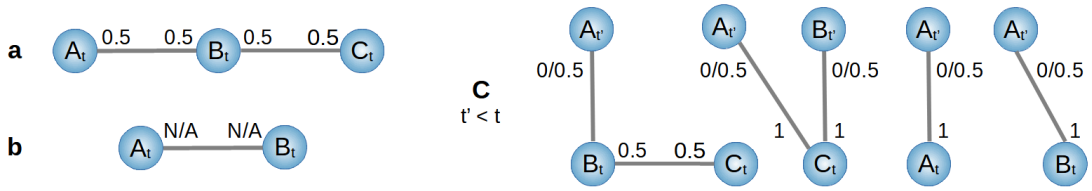


FIGURE 8.3 – **Probabilités initiales affectées aux arêtes. a** Pour les arêtes avec des nœuds contemporains dans un triplet ouvert, les extrémités sont initialisées avec une probabilité de 0.5. **b** Les arêtes avec des nœuds contemporains ne faisant pas partie d'un triplet ouvert ne peuvent pas être orientées. **c** Pour des arêtes avec des nœuds à des pas de temps différents, l'initialisation utilise le temps, y compris sans triplet ouvert. L'initialisation du côté le plus ancien des arêtes est différente si les variables latentes sont autorisées ($P_{init} = 0.5$) ou pas ($P_{init} = 0$).

L'algorithme non stationnaire envisagé est présenté à la page suivante.

Dans la mesure où cette version non stationnaire reste encore à mettre en œuvre, les benchmarks ne sont pas définis pour l'heure. Idéalement, les méthodes retenues pour réaliser l'évaluation devraient être capables de gérer des données mixtes, d'orienter les arcs contemporains et, si possible, d'autoriser les variables latentes.

L'application reste également à déterminer. Dans la mesure où le laboratoire fait partie de l'Institut Curie, l'application sera très certainement biologique avec, par exemple, l'étude d'impact d'un traitement ou la détermination de causes de l'apoptose (dans ce cas, en calant les trajectoires sur la mort de la cellule et s'intéressant aux effets antérieurs). La principale difficulté, ici, est de pouvoir disposer d'un ensemble d'expériences suffisant puisqu'il n'est pas possible de démultiplier les échantillons en les décalant dans le temps comme dans la version stationnaire.

Algorithm 6
Reconstruction causale par tMIIC non stationnaire (sans variable latente)

Require: τ the maximum time lag

- Preprocessing
Duplicate variables and data along τ back in history
- Skeleton reconstruction
$\mathcal{G} \leftarrow$ the complete graph on V
for all edges $X_t - Y_{t'} \in \mathcal{G}$ **do**
  if $\mathrm{I}'(X_t; Y_{t'}) \leqslant 0$ **then**
    Delete edge $X_t - Y_{t'}$ from $\mathcal{G}$
    Sepset$\{X_t, Y_{t'}\} \leftarrow \emptyset$
  else
    Find most contributing node $Z_{t''} \in \{\mathrm{adj}(X_t) \cup \mathrm{adj}(Y_{t'})\} \mid t'' \leq \max(t, t')$ which maximizes $R(X_t, Y_{t'}; Z_{t''} | \emptyset)$
  end if
end for
while There is a link $X_t - Y_{t'}$ with $R(X_t, Y_{t'}; Z_{t''} | \{U_i\}) > 1/2$ **do**
  for Top link $X_t - Y_{t'}$ with highest rank $R(X_t, Y_{t'}; Z_{t''} | \{U_i\})$ **do**
    Expand contributing set $\{U_i\} \leftarrow \{U_i\} + Z_{t''}$
    if $\mathrm{I}'(X_t; Y_{t'} | \{U_i\}) \leqslant 0$ **then**
      Delete edge $X_t - Y_{t'}$ from $\mathcal{G}$
      Sepset$\{X_t, Y_{t'}\} \leftarrow \{U_i\}$
    else
      Find next most contributing node $Z_{t''} \in \{\mathrm{adj}(X_t) \cup \mathrm{adj}(Y_{t'})\} \mid t'' \leq \max(t, t')$ and compute $R(X_t, Y_{t'}; Z_{t''} | \{U_i\})$
    end if
    Sort the rank list $R(X_t, Y_{t'}; Z_{t''} | \{U_i\})$
  end for
end while
- Skeleton orientation
Initialize $X_t - Y_{t'}$ as $X_t \to Y_{t'}$ if $t < t'$ and $X_t \leftarrow Y_{t'}$ if $t > t'$
Sort list of unshielded triples $\mathcal{L}_c = \{(X, Z, Y)_{X \neq Y}\}$ in decreasing order of $|\mathrm{I}'(X; Y; Z | \{U_i\})|$

... orientation from unshielded triples unchanged ...

return $\mathcal{G}$

8.2 Réseaux de régulation de gènes

Nous avons pu constater l'apport de traiter plusieurs milliers de gènes dans le cadre de la reconstruction de graphes de régulation des gènes en identifiant des facteurs de transcription que les méthodes classiques d'analyse d'expérimentations RNAseq en cellules uniques n'avaient pas repérés et ce domaine de recherche mériterait d'être investigué davantage.

A ce stade, les méthodes présentées dans ce mémoire de thèse utilisant l'a priori de conséquence nécessiteraient des développements complémentaires pour parvenir à devenir des outils utilisables facilement par la communauté scientifique.

En premier lieu, il serait nécessaire d'approfondir la méthode de sélection des facteurs de transcription et des autres gènes. Si nous avons déjà pu constater que la méthode traditionnellement utilisée de recherche de gènes marqueurs n'est pas adaptée dans notre cas, les avantages et inconvénients des sélections par changement de niveau d'expression ou via l'information mutuelle devront être évalués de façon plus approfondie. Il est également possible d'envisager d'autres manières de procéder, par exemple, si nous disposons de gènes d'intérêt tels les collagènes dans notre application, une possibilité serait de rechercher des facteurs de transcription spécifiquement enrichis par rapport aux gènes d'intérêt.

Sans préjuger des conclusions sur les méthodes de sélection, il est également possible que différentes méthodes soient retenues en fonction de l'application, par exemple, si le jeu de données contient ou pas des conditions expérimentales ou encore si une liste de gènes d'intérêt est fournie.

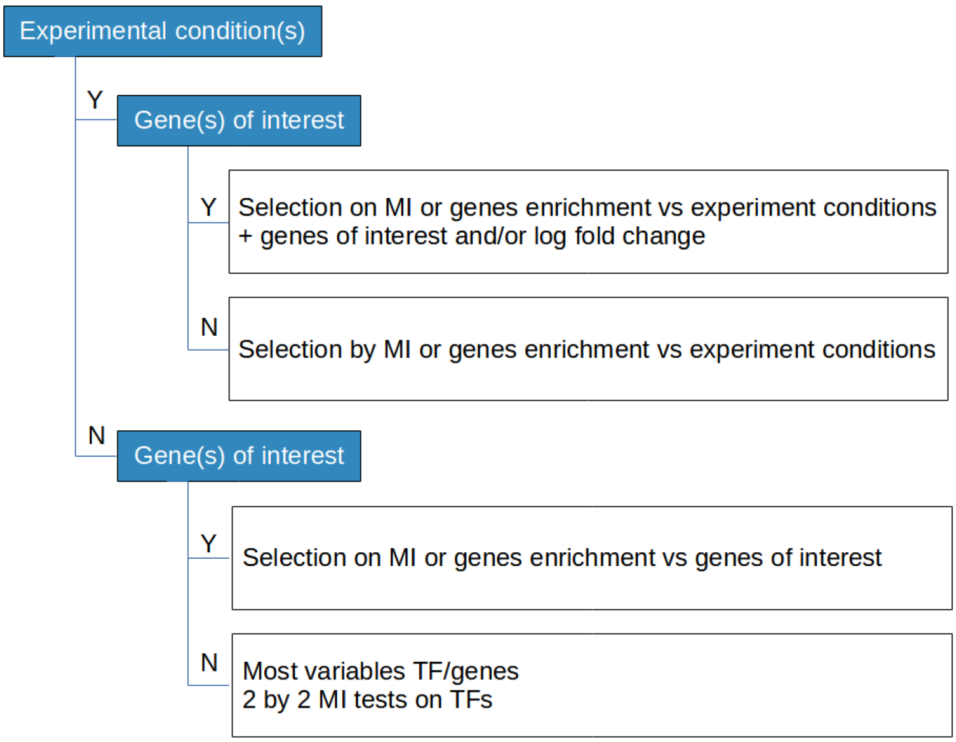


FIGURE 8.4 – **Quelques approches possibles pour développer tfMIIC.** Une version complète d'analyse de données scRNAseq devrait être applicable avec ou sans conditions expérimentales et avec ou sans gènes d'intérêt.

Plutôt que d'utiliser une condition expérimentale en entrée, il pourrait également être intéressant d'étendre la méthode pour comparer des types cellulaires. Sur l'application exposée dans le chapitre précédent, nous pourrions vouloir étudier les différences entre les cellules cancéreuses et les cellules cancéreuses inflammatoires afin d'identifier le réseau de gènes impliqués dans l'inflammation.

En imaginant qu'une telle méthode, qui pourrait être baptisée tfMIIC, soit pleinement implémentée, une étape de benchmarking sera bien sûr nécessaire. Parmi les méthodes existantes, des algorithmes tels que CLR, PCA-CMI, CMI2NI ou PCA-PMI [57], eux aussi basés sur l'information mutuelle, semblent de bons candidats ainsi que Genie3 qui est une des méthodes de référence dans le domaine.

Pour aller encore plus loin, d'autres extensions pourraient être envisagées comme l'intégration d'une dimension temporelle (comme dans DynGENIE3 ou BiXGBoost) ou la reconstruction de réseaux multi-niveaux. L'ajout de la dimension temporelle, notamment, pourrait être pour moi une direction naturelle de travaux futurs puisque cela constituerait une fusion de mes deux projets de recherche.

Conclusion

J'espère, au travers de ce document, avoir pu retranscrire combien, depuis maintenant plus de trois ans à l'institut Curie, j'ai pu acquérir une solide expérience en terme de démarche scientifique, de connaissances dans le domaine de la découverte causale ainsi que contribuer à la recherche en appliquant ces méthodes sur des jeux de données biologiques et par la mise à disposition de ces outils à l'ensemble de la communauté scientifique.

J'ai été ravi de pouvoir développer dans le cadre d'un mémoire les parties sur la théorie et les méthodes existantes ainsi que de pouvoir détailler mes travaux.

Ayant entrepris une reconversion il y a maintenant cinq ans pour intégrer la recherche dans le domaine de la santé humaine, un doctorat, outre la satisfaction personnelle, m'ouvrirait de nouvelles perspectives pour contribuer davantage à ce domaine, notamment en me permettant d'évoluer vers des postes d'ingénieur de recherche (IR) ou de post-doc, qui me sont actuellement inaccessibles sans doctorat.

Le fait d'intégrer la recherche en santé humaine me tenait à cœur et à présent, j'ai envie d'aller plus avant, d'approfondir et de développer mes travaux de recherche. Il est ainsi prévu qu'après le doctorat, si cette possibilité m'est offerte, je prenne un poste de post-doc pour continuer à développer mes travaux dans le domaine de la reconstruction de réseaux de gènes.

En vous remerciant du temps et de l'attention que vous porterez à ce mémoire, j'espère qu'il vous permettra de donner une issue favorable à ma démarche de doctorat.

Bibliographie

1. Pearl, J. *Causality : Models, Reasoning and Inference* 2nd. 478 p. (Cambridge University Press, USA, août 2009).
2. Glymour, M., Pearl, J. & Jewell, N. P. *Causal inference in statistics : A primer* (John Wiley & Sons, 2016).
3. Pearl, J. & Mackenzie, D. *The book of why : the new science of cause and effect* (Basic books, 2018).
4. Vladimir G. Ivancevic, T. T. I. *Computational Mind : A Complex Dynamics Perspective* (2007).
5. Hulswit, M. A Short History of 'Causation'. *SEED Journal (Semiotics, Evolution, Energy, and Development)* **4(3),** 16-42 (2004).
6. Falcon, A. in *The Stanford Encyclopedia of Philosophy* (éd. Zalta, E. N. & Nodelman, U.) Spring 2023 (Metaphysics Research Lab, Stanford University, 2023).
7. Wright, S. Correlation and Causation, 557-585. (1921).
8. Group, B. M. J. P. Streptomycin Treatment of Pulmonary Tuberculosis : A Medical Research Council Investigation. *Br Med J* **2.** Publisher : British Medical Journal Publishing Group Section : Article, 769-782 (30 oct. 1948).
9. Hannan, E. L. Randomized Clinical Trials and Observational Studies : Guidelines for Assessing Respective Strengths and Limitations. *JACC : Cardiovascular Interventions* **1,** 211-217 (1er juin 2008).
10. Assaad, C. *Causal Discovery between time series* thèse de doct. (Université Grenoble - Alpes, 5 juill. 2021).
11. Hitchcock, C. in *The Stanford Encyclopedia of Philosophy* (éd. Zalta, E. N.) Spring 2021 (Metaphysics Research Lab, Stanford University, 2021).
12. Höfler, M. Causal inference based on counterfactuals. *BMC Medical Research Methodology* **5,** 28 (13 sept. 2005).
13. Woodward, J. F. *Making things happen : a theory of causal explanation* (2023).
14. Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction, and Search* réd. par Berger, J. *et al.* (Springer, New York, NY, 1993).
15. Rubin, D. B. The design versus the analysis of observational studies for causal effects : parallels with the design of randomized trials. *Statistics in Medicine* **26,** 20-36 (2007).
16. ROSENBAUM, P. R. & RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41-55 (1er avr. 1983).
17. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38.** Number : 8 Publisher : Nature Publishing Group, 904-909 (août 2006).
18. Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V. & Klungel, O. H. Instrumental variables : application and limitations. *Epidemiology*, 260-267 (2006).

19. Assaad, C., Devijver, E. & Gaussier, E. Survey and Evaluation of Causal Discovery Methods for Time Series. *Journal of Artificial Intelligence Research* **73,** 767-819 (2022).

20. Verma, T. & Pearl, J. Equivalence and synthesis of causal models. Publisher : UCLA, Computer Science Department Los Angeles, CA (1991).

21. He, Y., Jia, J. & Yu, B. Counting and Exploring Sizes of Markov Equivalence Classes of Directed Acyclic Graphs.

22. Geiger, D. & Heckerman, D. in *Learning Gaussian Networks, Uncertainty Proceedings 1994* (éd. de Mantaras, R. L. & Poole, D.) 235-243 (Morgan Kaufmann, San Francisco (CA), 1er jan. 1994).

23. Heckerman, D., Meek, C. & Cooper, G. in *A Bayesian Approach to Causal Discovery, Innovations in Machine Learning : Theory and Applications* (éd. Holmes, D. E. & Jain, L. C.) 1-28 (Springer, Berlin, Heidelberg, 2006).

24. Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research* **3,** 507-554 (Nov 2002).

25. Peters, J. M. *Restricted structural equation models for causal inference* PhD Thesis (ETH Zurich, 2012).

26. Spirtes, P. & Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Social science computer review* **9,** 62-72 (1991).

27. Spirtes, P. *et al. Causation, prediction, and search* (MIT press, 2000).

28. Meek, C. *Causal inference and causal explanation with background knowledge* in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (Morgan Kaufmann Publishers Inc., 1995), 403-410.

29. Colombo, D. & Maathuis, M. H. Order-Independent Constraint-Based Causal Structure Learning. *J. Mach. Learn. Res.* **15,** 3741-3782 (2014).

30. Ramsey, J., Zhang, J. & Spirtes, P. L. *Adjacency-Faithfulness and Conservative Causal Inference* 27 juin 2012.

31. Li, H., Cabeli, V., Sella, N. & Isambert, H. Constraint-based Causal Structure Learning with Consistent Separating Sets (2019).

32. Shimizu, S., Hoyer, P. O., Hyvärinen, A. & Kerminen, A. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* **7** (2006).

33. Bühlmann, P., Peters, J. & Ernest, J. CAM : Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics* **42.** arXiv : 1310.1533, 2526-2556 (déc. 2014).

34. Affeldt, S. & Isambert, H. *Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information* in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence* (UAI, 2015), 42-51.

35. Affeldt, S., Verny, L. & Isambert, H. 3off2 : A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinformatics* **17,** 12 (2016).

36. Verny, L., Sella, N., Affeldt, S., Singh, P. P. & Isambert, H. Learning causal networks with latent variables from multivariate information in genomic data. *PLOS Computational Biology* **13,** 1-25 (2017).

37. Pearl, J. & Verma, T. S. in *Studies in Logic and the Foundations of Mathematics* 789-811 (Elsevier, 1995).

38. Cabeli, V. *et al.* Learning clinical networks from medical records based on information estimates in mixed-type data. *PLOS Computational Biology* **16.** Publisher : Public Library of Science, e1007866 (18 mai 2020).

39. Kontkanen, P. & Myllymäki, P. A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters* **103,** 227-233 (2007).

40. Roos, T., Silander, T., Kontkanen, P. & Myllymaki, P. *Bayesian network structure learning using factorized NML universal models* in *2008 Information Theory and Applications Workshop* (IEEE, 2008), 272-276.

41. Cover, T. M. & Thomas, J. A. Elements of information theory second edition solutions to problems. *Internet Access*, 19-20 (2006).

42. Howlader, N. *et al.* SEER cancer statistics review, 1975–2016. *National Cancer Institute* **1** (2019).

43. Gong, C., Yao, D., Zhang, C., Li, W. & Bi, J. *Causal Discovery from Temporal Data : An Overview and New Perspectives* 3 août 2023.

44. Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **37,** 424 (1969).

45. Schreiber, T. Measuring Information Transfer. *Physical Review Letters* **85,** 461-464 (2000).

46. Barnett, L., Barrett, A. B. & Seth, A. K. Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Physical Review Letters* **103** (2009).

47. Runge, J. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets.

48. Krich, C. *et al.* Decoupling between ecosystem photosynthesis and transpiration : A last resort against overheating. *Environmental Research Letters* **17** (1er avr. 2022).

49. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. & Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances* **5.** Publisher : American Association for the Advancement of Science, eaau4996 (27 nov. 2019).

50. Assaad, C. K., Devijver, E. & Gaussier, E. *Discovery of extended summary graphs in time series* in *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence* Uncertainty in Artificial Intelligence. ISSN : 2640-3498 (PMLR, 17 août 2022), 96-106.

51. Assaad, C. K., Devijver, E., Gaussier, E. & Ait-Bachir, A. *A Mixed Noise and Constraint-Based Approach to Causal Inference in Time Series* in *Machine Learning and Knowledge Discovery in Databases. Research Track* (éd. Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J. & Lozano, J. A.) (Springer International Publishing, Cham, 2021), 453-468.

52. Entner, D. & Hoyer, P. On Causal Discovery from Time Series Data using FCI. *Proceedings of the 5th European Workshop on Probabilistic Graphical Models, PGM 2010* (2010).

53. Malinsky, D. & Spirtes, P. *Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding* in *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery, CD@KDD 2018, London, UK, 20 August 2018* (éd. Le, T. D., Zhang, K., Kiciman, E., Hyvärinen, A. & Liu, L.) **92** (PMLR, 2018), 23-47.

54. Runge, J. *et al.* Inferring causation from time series in Earth system sciences. en. *Nat. Commun.* **10,** 2553 (2019).

55. Nguyen, M. *et al.* Dissecting Effects of Anti-cancer Drugs and Cancer-Associated Fibroblasts by On-Chip Reconstitution of Immunocompetent Tumor Microenvironments. *Cell Reports* **25,** 3884-3893.e3 (2018).

56. Munkres, J. Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math.* **5,** 32-38 (1957).

57. Zhao, M., He, W., Tang, J., Zou, Q. & Guo, F. A comprehensive overview and critical evaluation of gene regulatory network inference technologies. *Briefings in Bioinformatics* **22,** bbab009 (1er sept. 2021).

Publication principale (soumise)

CausalXtract: a flexible pipeline to extract causal effects from live-cell time-lapse imaging data

Franck Simon[1,‡], Maria Colomba Comes[2,‡], Tiziana Tocci[1,2,‡], Louise Dupuis[1], Vincent Cabeli[1], Nikita Lagrange[1], Arianna Mencattini[2], Maria Carla Parrini[3], Eugenio Martinelli[2,*], Hervé Isambert[1,*]

[1] CNRS UMR168, Institut Curie, Université PSL, Sorbonne Université, Paris, France

[2] Department of Electronic Engineering, University of Rome Tor Vergata, Rome, Italy

[3] INSERM U830, Institut Curie, Université PSL, Paris, France

[‡] these authors contributed equally to this work

[*] corresponding authors: herve.isambert@curie.fr, martinelli@ing.uniroma2.it

Live-cell microscopy routinely provides massive amount of time-lapse images of complex cellular systems. However, this wealth of data remains difficult to interpret in terms of causal effects. Here, we describe CausalXtract, a flexible computational pipeline that discovers causal and possibly time-lagged effects from morphodynamic features and cell-cell interactions in live-cell imaging data. We demonstrate the use of CausalXtract to uncover causal effects in a tumor-on-chip cellular ecosystem under therapeutically relevant conditions.

Live-cell imaging microscopy commonly produces extensive amounts of time-lapse images of cellular systems, which can be segmented to extract morphodynamic features and interactions of individual cells under increasingly complex and physiologically relevant conditions. However, this wealth of information remains largely under-exploited due to a lack of methods and tools able to discover causal effects from spatio-temporal correlations under well-controlled experimental conditions.

CausalXtract addresses this need by integrating an advanced live-cell image feature extraction module with a reliable and scalable causal discovery module, in order to learn temporal causal networks from live-cell time-lapse imaging data, Fig. 1.

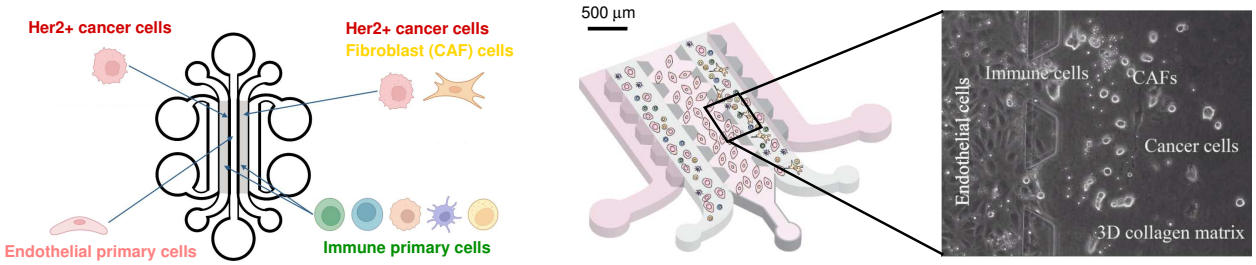
CausalXtract's live-cell image feature extraction module (CellHunter+), Fig. 1b, is based on CellHunter software[1] and consists in three steps: detection, tracking and feature extraction of live cells within time-lapse video images. First, automatic localization/segmentation of cells (*e.g.* tumor and immune cells) is performed with the Circular Hough Transform (CHT) algorithm[2] to estimate the cell centers and radii. Second, cell trajectories along the frames are constructed by linking the positions detected at the previous time step through Munkres' algorithm for Optimal sub-pattern Assignment Problems (OAPs)[3]. Finally, relevant descriptors related to the shape, motility, and state of the cells, as well as cell-cell interactions are quantified from each cell trajectory (Methods).

CausalXtract's temporal causal discovery module (tMIIC), Fig. 1c, is adapted from the causal discovery method, MIIC[4,5], which learns contemporaneous causal networks (*i.e.* when temporal information is not available) for a broad range of biological or biomedical data, from single-cell transcriptomic and genomic alteration data[4,6] to medical records of patients[5,7]. Live-cell time-lapse imaging data contain, however, information about cellular dynamics, which can in principle facilitate the discovery of novel cause-effect functional processes, based on the assumption that future events cannot cause past ones. To this end, CausalXtract's discovery module, tMIIC, reconstructs time-unfolded causal networks, where each variable is represented by several nodes at different relative time points[8], Fig. 1c. Such a time-unfolded network framework[9–12] is required to account for the temporal correlation between successive time steps in time series data. We benchmarked tMIIC on synthetic datasets resembling the real-world data of interest analyzed in this study (*i.e.* number of time steps, network size and degree distribution) and found that it matches or outperforms state-of-the-art methods, while running order of magnitudes faster on datasets of biologically relevant size including tens to hundreds of thousands time steps, Extended Data Figs. 1-4.

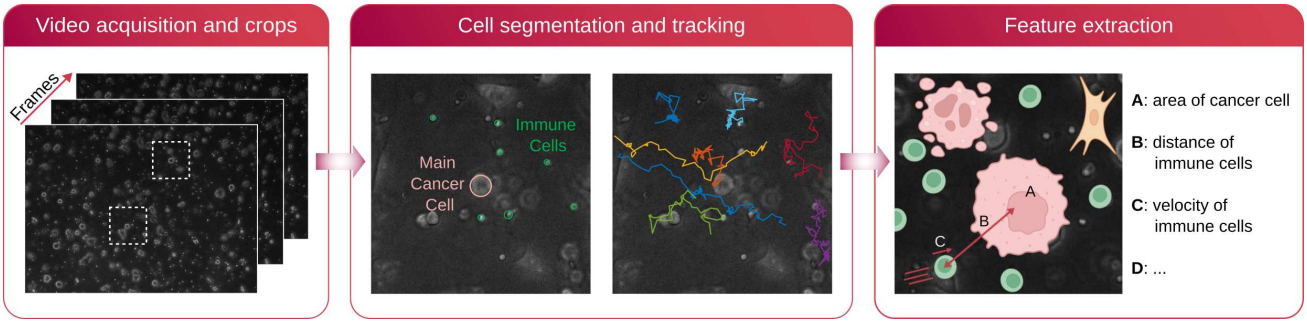
CausalXtract's temporal network framework goes beyond the seminal concept of temporal causality originally proposed by Granger[13] for linear time series without reference to graphical models and later extended to non-linear dynamics by Schreiber[14,15]. In particular, Granger-Schreiber causality is in fact too restrictive and may overlook actual causal effects, that can be uncovered by graph-based causal discovery methods, Extended Data Fig. 5 (Methods, Theorem 1). In addition, Granger-Schreiber causality has long been known to infer spurious causal associations based on time delays, by excluding the presence of latent common causes *a priori*[8]. CausalXtract circumvents these limitations by combining graph-based and information-based approaches (Methods), while including contemporary and time-delayed effects of unobserved latent variables, that are ubiquitous in cell biology data (*e.g.* the latent effects of cell cycle phases on cellular features and responses).

We showcase CausalXtract with the analysis of time-lapse images of a tumor ecosystem reconstituted *ex vivo* using the tumor-on-chip technology, Fig. 1a. These live-cell time-lapse images come from a proof-of-concept study[1] which demon-

a Tumor–on–chip preparation



b CausalXtract's live–cell image feature extraction module (CellHunter+)



c CausalXtract's temporal causal discovery module (tMIIC)

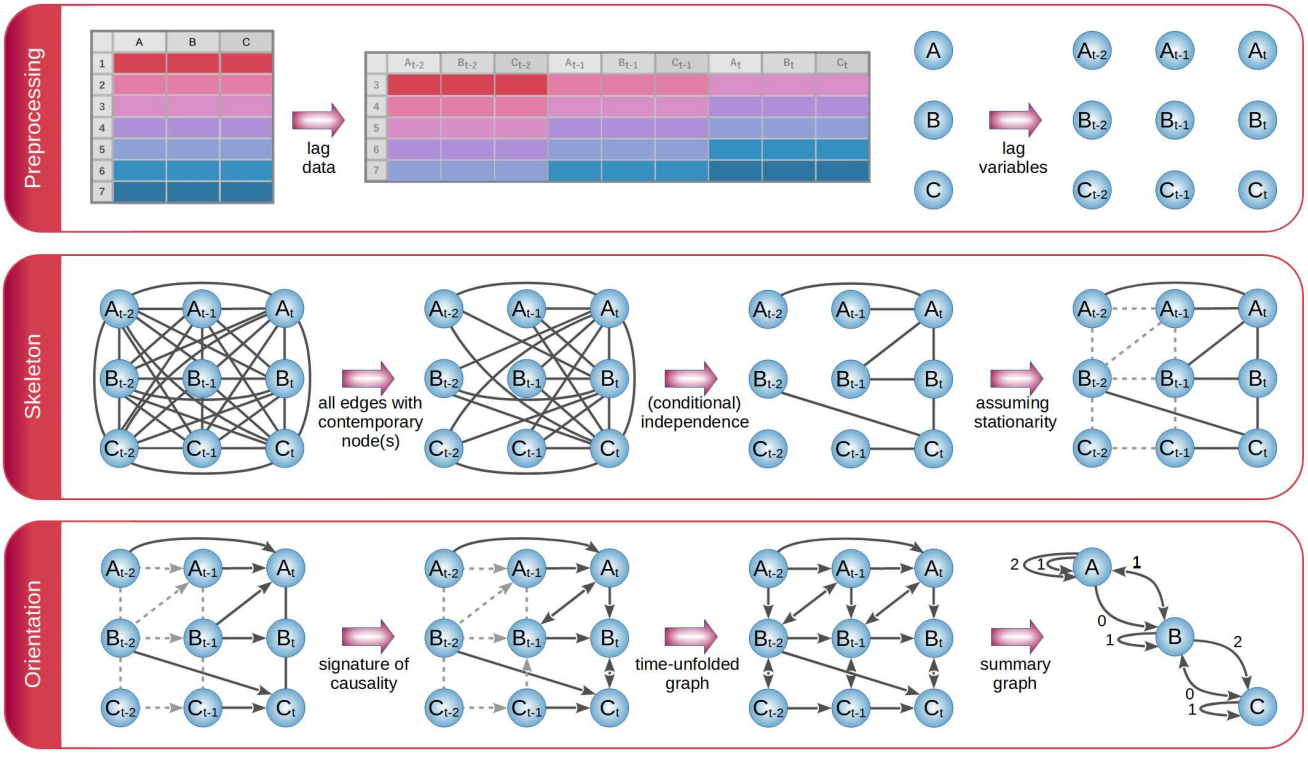


Figure 1: **CausalXtract pipeline.** **a**, Live-cell tumor ecosystem reconstituted *ex vivo*[1] using the tumor-on-chip technology (Methods). **b**, CausalXtract's live-cell image feature extraction module (CellHunter+). The tracking of cancer and immune cells and of their mutual interactions is illustrated in Supplementary Movies 1-3, in absence or presence of cell division and apoptosis event. **c**, CausalXtract's temporal causal discovery module (tMIIC) learns a temporal causal network from the features extracted in (b). See Methods for CausalXtract's implementation details and theoretical foundations. A step-by-step notebook of CausalXtract pipeline is provided with the source code.

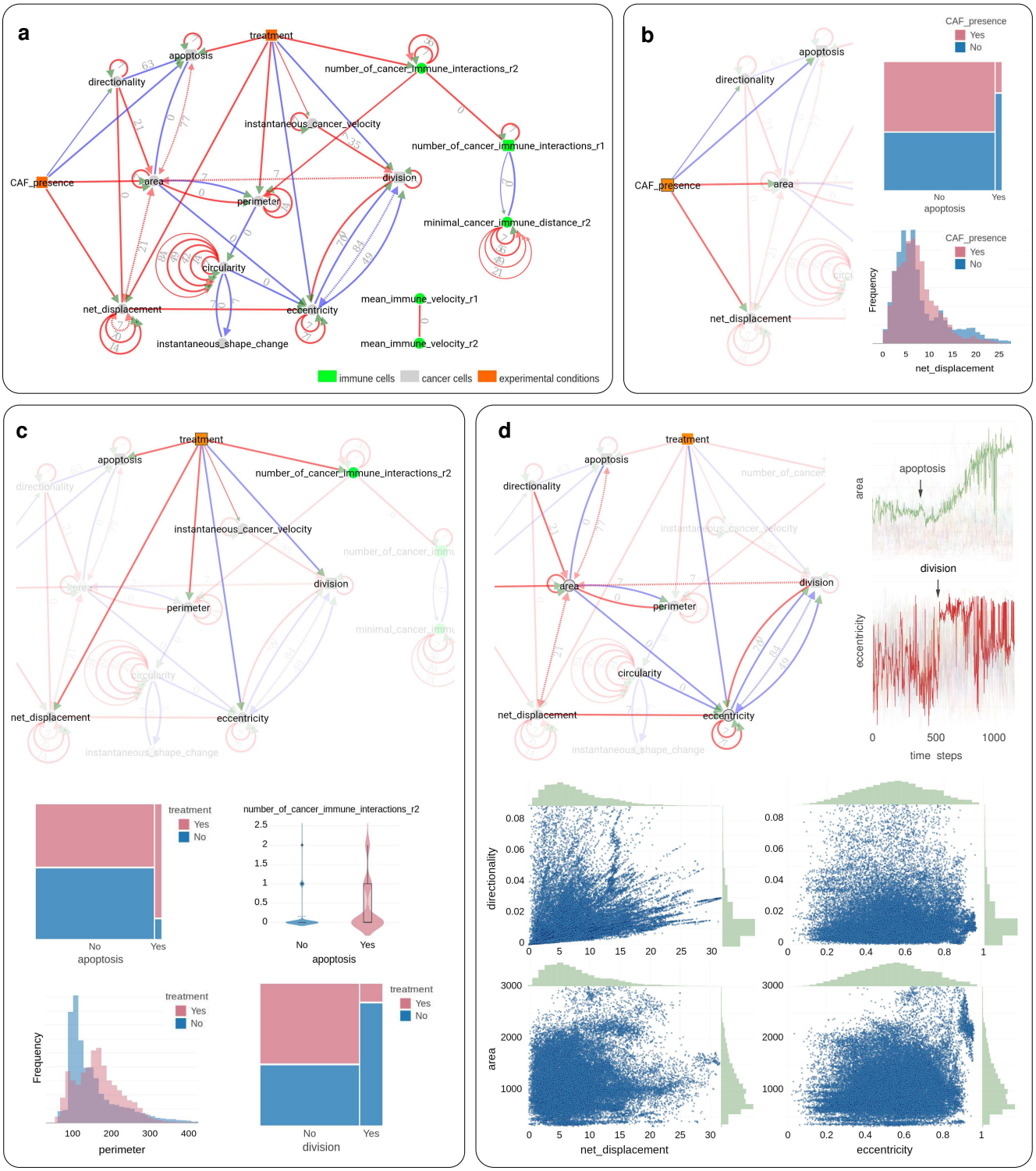


Figure 2: **Application of CausalXtract to time-lapse images of tumor ecosystems reconstituted ex vivo[1].** **a**, Summary causal network inferred by CausalXtract. The underlying time-unfolded causal network is shown on Extended Data Fig. 7. Red (resp. blue) edges correspond to positive (resp. negative) associations. Bidirected dashed edges represent the effect of unobserved (latent) common causes. Annotations on edges correspond to time delays in time-steps (1 ts = 2 min). The inferred network is largely robust to variations in sampling rate ($\delta\tau$) and maximum lag (τ), Extended Data Fig. 8. Here $\delta\tau = 7$ ts and $\tau = 84$ ts are chosen automatically by CausalXtract, Extended Data Fig. 8b. **b**, The CAF presence subnetwork highlighting the direct causal effects of CAFs on cancer cells. In particular, CausalXtract uncovers that CAFs directly inhibit cancer cell apoptosis independently from treatment, which has not been reported so far. **c**, The treatment subnetwork highlighting the direct causal effects of treament on cancer cells. In particular, CausalXtract uncovers that treatment increases cancer cell perimeter, which has not been reported either. **d**, The eccentricity-area subnetwork highlighting multiple direct and possibly antagonistic time-lagged effects, notably, between cell division and eccentricity and between cell apoptosis and area, as discussed in main text.

strated the effects of an anti-cancer drug (the monoclonal antibodies trastuzumab, brand name Herceptin, used to treat HER2+ breast cancers) on a reconstituted tumor microenvironment including cancer cells, immune cells, cancer-associated fibroblasts (CAF), and endothelial cells (Methods). However, a comprehensive extraction and analysis of cellular morphodynamic features and interactions remained unexplored.

To this end, cellular features such as cell geometry, velocity, division, apoptosis, cell-cell transient interactions and persistent contacts were first extracted from the raw images using CausalXtract's feature extraction module, Fig. 1b and Extended Data Fig. 6. Then, a time-unfolded causal network, Extended Data Fig. 7, and the corresponding summary causal network, Fig. 2a, were reconstructed between extracted cellular features, cell-cell interactions and therapeutic conditions using CausalXtract's temporal causal discovery module, Fig. 1c.

CausalXtract inferred network, Fig. 2a, uncovers novel biologically relevant findings, in addition to confirming known results from earlier studies. In particular, CausalXtract discovers that CAFs directly inhibit cancer cell apoptosis, independently from anti-cancer treatment, Fig. 2b, while earlier studies reported that CAFs merely reduced the effect of treatment[1]. CausalXtract also discovers that treatment increases cancer cell perimeter, Fig. 2c, which has not been reported so far either. In addition, CausalXtract confirms known results from earlier studies. In particular, it recovers that treatment increases cancer cell apoptosis and the number of cancer-immune interactions, as well as decreases the division rate of cancer cells, Fig. 2c. Likewise, CausalXtract recovers that CAFs stimulate cancer cell migration and increase their area, Fig. 2b.

Interestingly, CausalXtract identifies also multiple and possibly antagonistic effects with different time delays. For instance, CausalXtract recovers several antagonistic relations between morphodynamic features such as cell division and eccentricity or cell apoptosis and area, Fig. 2d. Indeed, the late phases of cell division are associated to a marked increase in eccentricity (red edge) but preceded by a net decrease in eccentricity, two to three hours before cytokinesis (blue edges), once the decision to divide has been made (*i.e.* the probable latent cause) and the cell is actually duplicating its biological materials (prophase), Fig. 2d. Likewise, the area change upon apoptosis is predicted to first decrease soon after apoptosis (blue edge) before eventually increasing upon cell lysis (red edge), Fig. 2d. These results are robust to variations in sampling rate, Extended Data Fig. 8.

All in all, CausalXtract is a flexible pipeline which uncovers novel and possibly time-lagged causal relations between cellular features under controlled conditions (*e.g.* drug). CausalXtract consists of two independent modules, conceived to warrant interoperability with alternative live-cell segmentation and tracking methods or alternative temporal causal discovery methods.

CausalXtract opens up new avenues to analyze live-cell imaging data for a range a fundamental and translational research applications, such as the use of tumor-on-chips to screen immunotherapy responses on patient-derived tumor samples. With the advent of virtually unlimited live-cell image data, flexible hypothesis-free interpretation pipelines are much needed and we believe that CausalXtract can bring unique insights based on causal discovery to interpret such underexploited live-cell imaging data.

References

1. Nguyen, M. *et al.* Dissecting Effects of Anti-cancer Drugs and Cancer-Associated Fibroblasts by On-Chip Reconstitution of Immunocompetent Tumor Microenvironments. *Cell Reports* **25,** 3884–3893.e3 (2018).
2. Davies, E. *Machine vision* 3rd ed. (Morgan Kaufmann, 2004).
3. Munkres, J. Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math.* **5,** 32–38 (1957).
4. Verny, L., Sella, N., Affeldt, S., Singh, P. P. & Isambert, H. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.* **13,** e1005662 (2017).
5. Cabeli, V. *et al.* Learning clinical networks from medical records based on information estimates in mixed-type data. *PLoS Comput. Biol.* **16,** e1007866 (2020).
6. Desterke, C. *et al.* Inferring Gene Networks in Bone Marrow Hematopoietic Stem Cell-Supporting Stromal Niche Populations. *iScience* **23,** 101222 (2020).
7. Sella, N. *et al.* Interactive exploration of a global clinical network from a large breast cancer cohort. *npj Digital Med* **5,** 113 (2022).
8. Assaad, C., Devijver, E. & Gaussier, E. Survey and Evaluation of Causal Discovery Methods for Time Series. *Journal of Artificial Intelligence Research* **73,** 767–819 (2022).
9. Entner, D. & Hoyer, P. On Causal Discovery from Time Series Data using FCI. *Proceedings of the 5th European Workshop on Probabilistic Graphical Models, PGM 2010* (2010).
10. Malinsky, D. & Spirtes, P. *Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding* in *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery, CD@KDD 2018, London, UK, 20 August 2018* (eds Le, T. D., Zhang, K., Kiciman, E., Hyvärinen, A. & Liu, L.) **92** (PMLR, 2018), 23–47.
11. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. & Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances* **5** (2019).
12. Runge, J. *et al.* Inferring causation from time series in Earth system sciences. en. *Nat. Commun.* **10,** 2553 (2019).
13. Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **37,** 424 (1969).
14. Schreiber, T. Measuring Information Transfer. *Physical Review Letters* **85,** 461–464 (2000).
15. Barnett, L., Barrett, A. B. & Seth, A. K. Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Physical Review Letters* **103** (2009).
16. Chan, T. F. & Vese, L. A. Active contours without edges. *IEEE Trans. Image Process.* **10,** 266–277 (2001).
17. Masuzzo, P., Van Troys, M., Ampe, C. & Martens, L. Taking aim at moving targets in computational cell migration. *Trends Cell Biol.* **26,** 88–110 (2016).
18. Pearl, J. *Causality* (Cambridge university press, 2009).
19. Spirtes, P., Glymour, C. N., Scheines, R. & Heckerman, D. *Causation, prediction, and search* (MIT press, 2000).
20. Runge, J. *Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets* in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)* (eds Peters, J. & Sontag, D.) **124** (PMLR, 2020), 1388–1397.

Methods

Tumor-on-chip preparation and live-cell microscopy. Videos analyzed in the present study refer to biological experiments emulating a 3D breast tumor ecosystem[1]. All tumor-on-chip experiments have a central endothelium compartment containing endothelial cells (primary human umbilical vein endothelial cells, HUVECs) and two lateral chambers filled with biomimetic hydrogel (collagen type I at 2.3 mg/mL) seeded with cancer cells (HER2+ breast cancer BT474 cell line) and immune cells (peripheral blood mononuclear cells, PBMCs) from healthy donors, Fig. 1a. Four experimental conditions were considered depending on the presence or absence of breast cancer-associated fibroblasts (CAF cell line Hs578T) and drug treatment (trastuzumab, Herceptin). Videos were acquired by inverted motorized Leica microscopes with a frame rate of 2 minutes for up to 48h (1440 frames). Fig. 1b shows a crop frame with cancer cells, PBMCs and CAFs. Each video was cropped into multiple small 300×300 pixel videos (referred to as crops in the following), each of which represented a field of view at subsequent time frames containing a "main" cancer cell (MCC) initially placed at the center of the image, some PBMC immune cells, other cancer cells and possibly CAFs within the surrounding of the MCC depending on the experimental conditions. 36 video crops of up to 1440 frames were analyzed (46,935 frames in total) corresponding to 9 video crops per experimental conditions.

CausalXtract's live-cell image feature extraction module. The live-cell image feature extraction module (CellHunter+), Fig. 1b, extends the CellHunter software[1] and consists in three steps: detection, tracking and feature extraction of live cells within time-lapse video images. First, cell detection is based on the segmentation of circular-shaped objects using CHT[2] with radii set around the theoretical radii of the two cell populations ($r_{\mathrm{im}} = 4\,\mathrm{px}$ for immune cells and $r_{\mathrm{ca}} = 14\,\mathrm{px}$ for MCCs with a pixel resolution $1\,\mathrm{px} = 0.645\,\mu\mathrm{m}$[1]). Then, cell tracking is performed by linking cells detected at the i^{th} frame to cells located at the $(i+1)^{th}$ frame within a maximum distance from the detected cell candidate. While the motions of both MCCs and immune cells ressemble random walks with time-varying drift and volatility, these two cell types exhibit different motility characteristics[1]. Hence, different maximum distances are considered for the two cell populations: it was set to 40 px for MCCs and to 20 px for immune cells. For each cell population, an OAP using the Munkres algorithm[3] is solved: the globally best possible pairing among located objects is based on an assignment cost equal to the inverse of the distance between pairs of cell candidates at the i^{th} and $(i+1)^{th}$ frames. Cell appearing/disappearing and cell overlaps due to projection errors of the 3D scene in the 2D domain are also handled. Finally, cellular morphodynamic features and cell-cell interaction features are extracted at successive positions along each trajectory. For each MCC, 15 descriptors were extracted, Extended Data Fig. 6, and classified into four main categories: cell shape, motility, state, and interaction descriptors.

Shape descriptors. The active contour algorithm implemented in Matlab[16] was used to segment the MCC boundaries on each video crop frame. Taking as input a frame representing the i^{th} snapshot of the t^{th} MCC, it returns a binary image, where the MCC is represented by a white region. From the binary image, the shape properties of the region occupied by each MCC were extracted using the Matlab *regionprops* algorithm. The resulting descriptors of the extracted shape are listed below:

- *area* indicates the number of pixels composing the region. The equivalent diameter of the t^{th} MCC in the i^{th} frame is defined as $d_i^t = \sqrt{4 \cdot area/\pi}$.
- *perimeter* represents the distance along the MCC boundary.
- *circularity* is defined as $4 \cdot area \cdot \pi / perimeter^2$, which is equal to 1 when the region is perfectly circular.
- *eccentricity* denotes the eccentricity of the ellipse with the same second moments as the region. The value is equal to 1 when the region is a line and to 0 when the region is a circle.
- *instantaneous shape change* is defined as, $|d_i^t - d_{i-1}^t|$, corresponding to the difference in absolute value of the equivalent diameters between the i^{th} and $(i-1)^{th}$ frames of the t^{th} MCC.

Motility descriptors. The positions $p_i^t = (x_i^t, y_i^t)$ and p_{i-1}^t of the t^{th} MCC in the i^{th} and $(i-1)^{th}$ frames were compared using the Euclidean distance $d(\cdot)$ to define the following motility parameters:

- *instantaneous cancer velocity*[17] is defined as $d(p_i^t, p_{i-1}^t)/\Delta t$, where Δt is the time interval between two consecutive frames.
- *net displacement*[17] indicates the resultant distance between the initial and current positions of the t^{th} MCC, $d(p_1^t, p_i^t)$.
- *directionality*[17] is defined as the ratio of net displacement, $d(p_1^t, p_i^t)$, and curvilinear distance, $\sum_{k=2}^{i} d(p_k^t, p_{k-1}^t)$. It measures the persistence of motion and ranges from 0 for confined cells to 1 for cells moving perfectly straight in one direction.

State descriptors. They record apoptosis or division events:

- *apoptosis* indicates if the MCC has died during the experiment. It is set to 'No' as long as the cell has not died and becomes 'Yes' for the remaining frames after the cell undergoes apoptosis.
- *division* indicates if the MCC has divided during the experiment. It is set to 'No' as long as the cell has not divided and becomes 'Yes' for the remaining frames after the cell divides.

Interaction descriptors. Interactions between MCCs and immune cells were defined with respect to two radii around each MCC, $r_1 = r_{\mathrm{im}} + r_{\mathrm{ca}} + 2 = 20\,\mathrm{px}$ and $r_2 = 2 \times (r_{\mathrm{im}} + r_{\mathrm{ca}}) = 36\,\mathrm{px}$[1]. Hence, r_1 refers to MCC and immune cells in actual physical contact, while r_2 refers to MCC and immune cells in close vicinity. Then, for each sample the following interaction features were defined:

- *number of cancer-immune interactions (r_2)* corresponds to the number of immune cells within the interaction radius r_2 around the MCC on that frame.
- *number of cancer-immune interactions (r_1)* corresponds to the number of immune cells in close contact with the MCC on that frame.
- *minimal cancer-immune distance (r_2)* is the minimum distance between the MCC and the immune cells within a radius r_2.
- *mean immune velocity (r_2)* is the mean instantaneous velocity norm of the immune cells within the interaction radius r_2 around the MCC.
- *mean immune velocity (r_1)* is the mean instantaneous velocity norm of the immune cells in close contact with the MCC.

Overview of causal discovery methods for non-temporal data. Traditional causal discovery methods[18,19] aim to learn causal networks from datasets of independent samples by proceeding through successive steps. They first learn structural constraints in the form of unconditional or conditional independence between variables and remove the corresponding edges from an initial fully connected network. The second step then consists in orienting some of the retained edges based on the signature of causality in observational data. This corresponds to orienting three-variable "v-structure" motifs as, $X \rightarrow Z \leftarrow Y$, whenever the edge $X - Y$ has been removed without conditioning on the variable Z, which implies that Z cannot be a cause of X nor Y. This does not guarantee, however, that X (or Y) is an actual cause of Z, which also requires to rule out the possibility that the edge between X and Z (or Y and Z) might originate from a latent common cause, L, unobserved in the dataset, *i.e.* $X \dashleftarrow L \dashrightarrow Z$. The recent causal discovery method, MIIC[4,5], learns more robust causal graphical models by first collecting iteratively significant information contributors before assessing conditional independences based on a mutual information supremum principle for finite dataset[5]. In practice, MIIC's strategy limits spurious conditional independences which improves its

edge sensitivity and orientation reliability compared to traditional constraint-based methods[4,5]. In addition, MIIC can handle heterogeneous data[5] (*i.e.* combining continuous and categorical variables), missing data[5] and unobserved latent variables[4], that are ubiquitous in real-world applications.

CausalXtract's causal discovery module for time series data In order to analyze time series datasets, CausalXtract's causal discovery module (tMIIC) aims to learn a time-unfolded graph, $\mathcal{G}_t$, where each variable is represented by a series of nodes associated to its value at different relative time points, Fig. 1c. Such a time-unfolded network framework[9–12] is required to account for the temporal correlation between successive samples in time series data. Assuming that the dynamics can be considered stationary (see Benchmarking of CausalXtract's causal discovery module section, below), the time-unfolded graph, $\mathcal{G}_t$, should be translationally invariant over time and can be assigned a periodic structure *a priori*. In addition, $\mathcal{G}_t$ can be restricted to a few time steps from the running time, t, back to a maximum time lag, $t-\tau$, since nodes at future time points ($t' > t$) cannot *a priori* influence the observed data at current or previous time points ($t' \leqslant t$), Fig. 1c. The maximum time lag τ should be chosen so as to have little effect on the final graphical model, which can be achieved for instance by setting τ to twice the relaxation time of the slowest variables of the dataset. In practice, we may also limit the number of time points ν in $\mathcal{G}_t$ by introducing a time increment $\delta\tau$ between consecutive time points, which leads to $\nu = \tau/\delta\tau$ time-lagged layers in $\mathcal{G}_t$. Such a compact periodic graphical representation over a sliding temporal window can be efficiently learned by adapting MIIC causal discovery method for non-temporal data to identify all necessary edges involving at least one contemporaneous node at time t, Fig. 1c. Once these time-lagged and contemporaneous edges have been identified, they are simply duplicated at earlier time points to enforce the translational invariance of $\mathcal{G}_t$ skeleton. Time-lagged edges are then pre-oriented with a first arrowhead pointing towards the future, considering that current time points cannot cause earlier events. Then, contemporaneous and time-lagged edges can be further oriented using MIIC orientation probability scores applied to $\mathcal{G}_t$, which may also uncover a second arrowhead (backward in time) for time-lagged edges. This corresponds to time-lagged latent causal effects from unobserved common causes, Fig. 1c.

Learning such structural models including latent variables from time series data was first proposed for time-lagged effects[9] and subsequently extended to contemporaneous effects[10] by adapting the constraint-based FCI method allowing for latent variables[19]. While traditional constraint-based methods suffer from poor recall, the recent PCMCI[11] / PCMCI+[20] method improves recall by introducing ad hoc conditioning rules for auto-correlated time series. By contrast, tMIIC does not require any ad hoc conditioning rules, as it relies on the same information-theoretic strategy as MIIC to limit spurious independence and improve edge recall. tMIIC also captures time-lagged and contemporaneous effects due to latent variables.

Relation to Granger-Schreiber temporal causality. The concept of temporal causality was originally formulated by Granger[13] without reference to any graphical model by comparing linear autoregression with or without past values of possible causal variables. This was later extended to non-linear relations by Schreiber[14,15] using the notion of Transfer Entropy, $T_{X\to Y}$, which can be expressed in terms of multivariate conditional information,

$$T_{X\to Y} = I(Y_t; \boldsymbol{X}_{t'<t} | \boldsymbol{Y}_{t'<t}) \tag{1}$$

where $\boldsymbol{X}_{t'<t}$ and $\boldsymbol{Y}_{t'<t}$ denote the sets of variables, $X_{t'}$ and $Y_{t'}$, taken at earlier time points t' than t.
While Eq. 1 is asymmetric upon X/Y permutation, a simple comparison of Transfer Entropy asymmetry (*e.g.* $T_{X\to Y} > T_{Y\to X} \geqslant 0$) does not necessarily translate into causal direction as this asymmetry is also expected for non-causal relations. Interestingly, this is in fact the absence of Transfer Entropy in one direction (*e.g.* $T_{Z\to X} \approx 0$) which suggests the possibility of a causal relation in the opposite direction, $X \to Z$, as in the case of v-structures in graph-based causal discovery methods, provided that a latent common cause can be excluded between the two variables (as discussed above).

We clarify in Theorem 1 below this relation between temporal causality without reference to any structural model (Eq. 1) and structural causality entailed by time-unfolded causal graphical models ($\mathcal{G}_t$). This highlights the common foundations of temporal and structural causalities beyond their seemingly unrelated definitions.

Theorem 1. [$T_{Y\to X} = 0$ implies temporal ($2\ var + t$) v-structures] If X_t is adjacent to Y_t in $\mathcal{G}_t$ and $T_{Y\to X} = I(X_t; \boldsymbol{Y}_{t'<t} | \boldsymbol{X}_{t'<t}) = 0$, then for all $Y_{t'}$ adjacent to Y_t in $\mathcal{G}_t$, with $t' < t$, there is a temporal (2 variable + time) v-structure, $Y_{t'} \to Y_t \leftarrow X_t$, in $\mathcal{G}_t$.

Proof: We reason by contradiction based on Extended Data Fig. 5a: if there exists $Y_{t'}$ adjacent to Y_t such that $Y_{t'} - Y_t - X_t$ is not a v-structure, then $T_{Y\to X} = I(X_t; \boldsymbol{Y}_{t'<t} | \boldsymbol{X}_{t'<t}) \neq 0$, as $Y_t \notin \boldsymbol{X}_{t'<t}$ or X_t is adjacent to $Y_{t'}$. □

Note, however, that the converse of Theorem 1 is not true: a temporal v-structure does not imply a vanishing Transfer Entropy, as shown with the counterexample in Extended Data Fig. 5b. As a result, the presence of a temporal v-structure, $Y_{t'} \to Y_t \leftarrow X_t$ in $\mathcal{G}_t$, does not necessarily imply a vanishing transfer entropy, $T_{Y\to X} = 0$, as long as there remains an edge between any $Y_{t''}$ and X_t, as in the example in Extended Data Fig. 5b. Hence, Granger-Schreiber causality is in fact too restrictive and may miss actual causal effects, which can be uncovered by structural causal discovery methods like tMIIC. In addition, Granger-Schreiber causality is also known to infer spurious causal associations by excluding the presence of latent common causes *a priori*. By constrast, CausalXtract's causal discovery module includes time-delayed as well as synchronous effects originating from unobserved latent variables, as discussed above.

Benchmarking of CausalXtract's causal discovery module. The performance of CausalXtract's causal discovery module (tMIIC) has been assessed using Tigramite package[20], which provides different methods to learn temporal causal networks from time series data. We compared tMIIC to two methods capable of orienting contemporaneous edges (PC and PCMCI+) and tested three different kernels for estimating mutual information (Parcorr, GPDC and KNN). Benchmark networks and datasets have been chosen to resemble the real-world data analyzed in this study (*i.e.* similar number of time steps, network size and degree distribution) and include a large range of linear and non-linear relations between variables.

A first series of datasets was generated for a 15 node benchmark network (Extended Data Fig. 1a) with linear combinations of contributions inspired by the Tigramite package, Supplementary Table 1. Running times and scores (Precision, Recall, F-score) have been averaged over 10 datasets (Extended Data Fig. 1b) and show that tMIIC scores are at par with PC and PCMCI+ using GPDC or KNN kernels but that tMIIC runs orders of magnitude faster, which enables to use tMIIC on much larger datasets of biological interest including a few tens or hundreds of thousands samples. Only PC or PCMCI+ using ParCorr kernel match tMIIC running speed but with significantly lower scores, as Fscores level off around 0.6-0.7 at large sample size, while tMIIC Fscore exceeds 0.9 (Extended Data Fig. 1b).

Importantly, increasing the number of time-lagged layers from $\tau = 2$ (as in the actual model, Extended Data Fig. 1a) to 5 or 10 layers in the inferred time-unfolded network (Extended Data Fig. 2) leads to very similar network reconstructions for simulated stationary data. This demonstrates tMIIC insensitivity to an overestimated maximum lag for the reconstituted network. Interestingly, however, when the generated data is no longer stationary, increasing the number of layers leads to multiple self-loops at non-stationary variables, whilst the rest of the network remains relatively unaffected (Extended Data Fig. 3). It demonstrates that CausalXtract's causal discovery module is robust to the presence of non-stationary variables but requires long-time range interactions, and therefore multiple time-lagged layers, to account for these non-stationary dynamics

at specific variables. This striking observation on benchmark networks is also consistent with the multiple self-loops observed for a number of non-stationary variables in the real-world application on cellular ecosystems, Fig. 2a and Extended Data Fig. 6.

A second series of more complex datasets was also generated for another 15 node benchmark network (Extended Data Fig. 4a) with non-linear combinations of contributors, Supplementary Table 2. Here, tMIIC tends to outperform both PC and PCMCI+, in terms of Recall and Fscores, while remaining orders of magnitude faster compared to GPDC and KNN kernels. Only PC or PCMCI+ using ParCorr kernel match tMIIC running speed but with significantly lower scores (*i.e.* Fscores level off around 0.4-0.5 at large sample size, while tMIIC Fscore exceeds 0.8). This demonstrates that CausalXtract's causal discovery module (tMIIC) is both a reliable and scalable method to discover complex temporal causal relations in very large time series datasets including a few hundred thousand samples.

Data availability

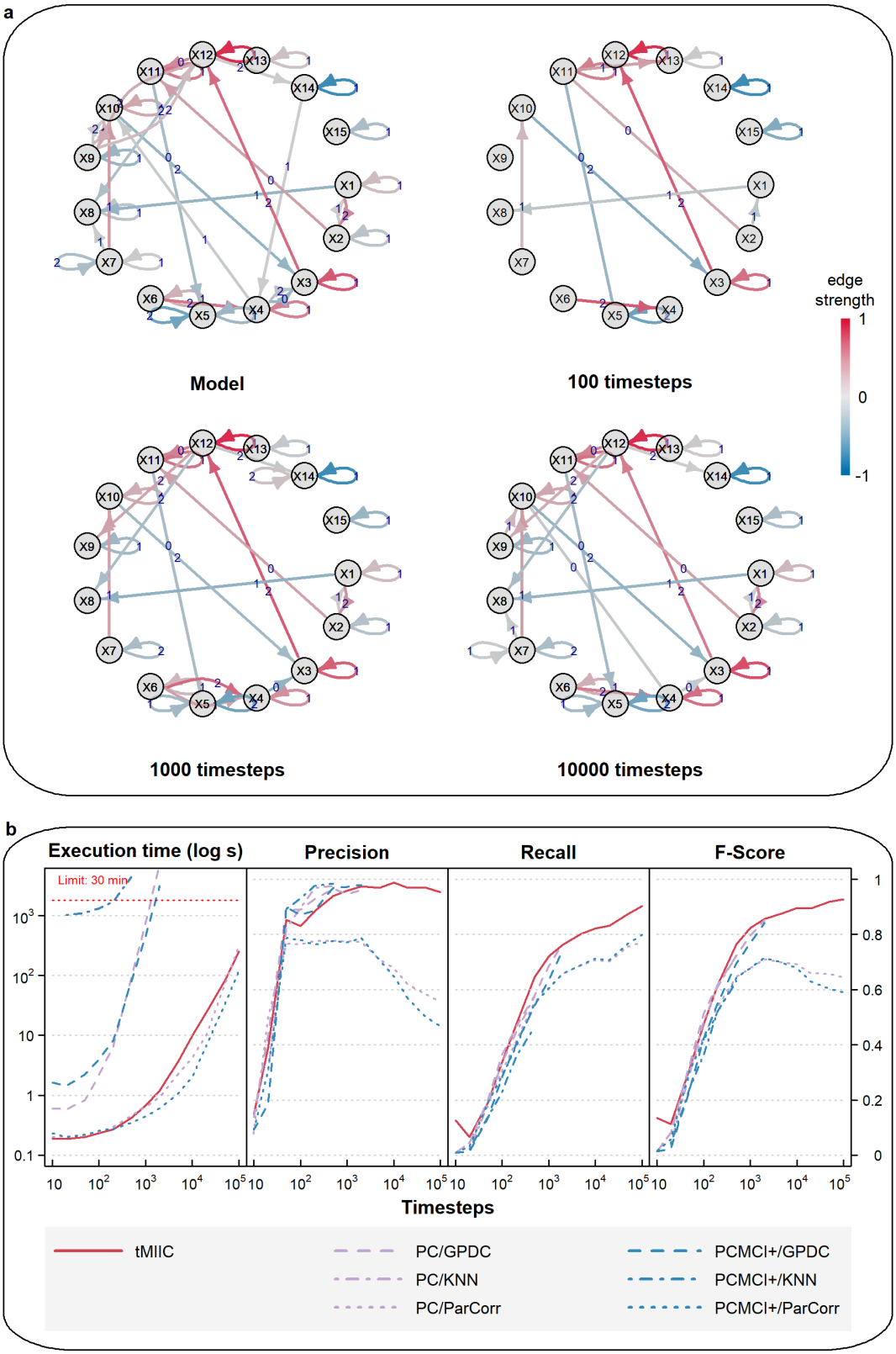
The original live-cell time-lapse image data and extracted crops are available at: `https://doi.org/10.5281/zenodo.7755699`.

Code availability

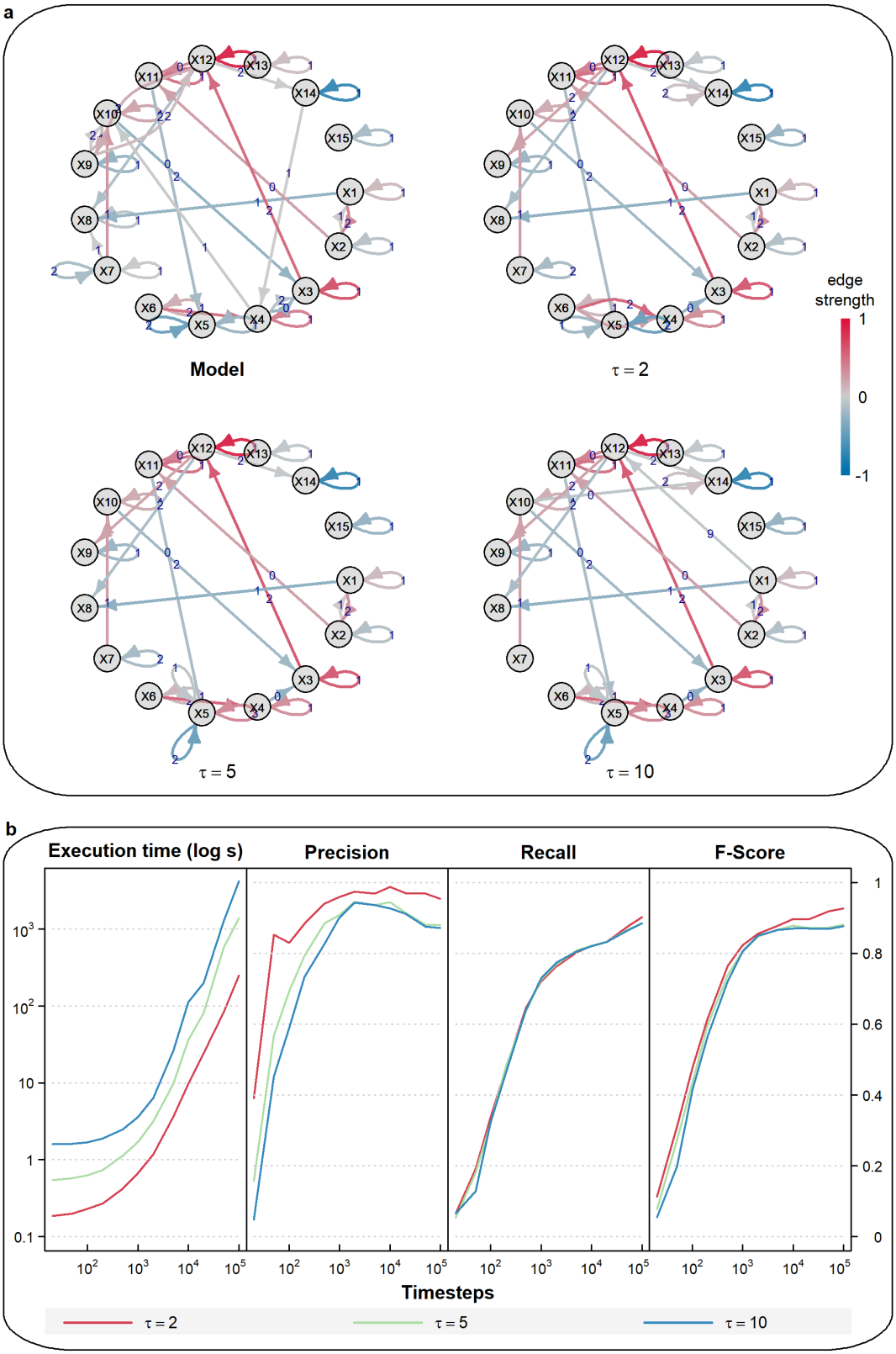
The source code of the CausalXtract pipeline is available at: `https://github.com/miicTeam/CausalXtract`. It includes a demo R markdown notebook of CausalXtract pipeline, which reproduces step-by-step the results reported in the manuscript, Fig. 2, starting from the original live-cell time-lapse images of the tumor-on-chip ecosystem, Fig. 1a. Tigramite package used for benchmark comparison is available at: `https://github.com/jakobrunge/tigramite`

Acknowledgements

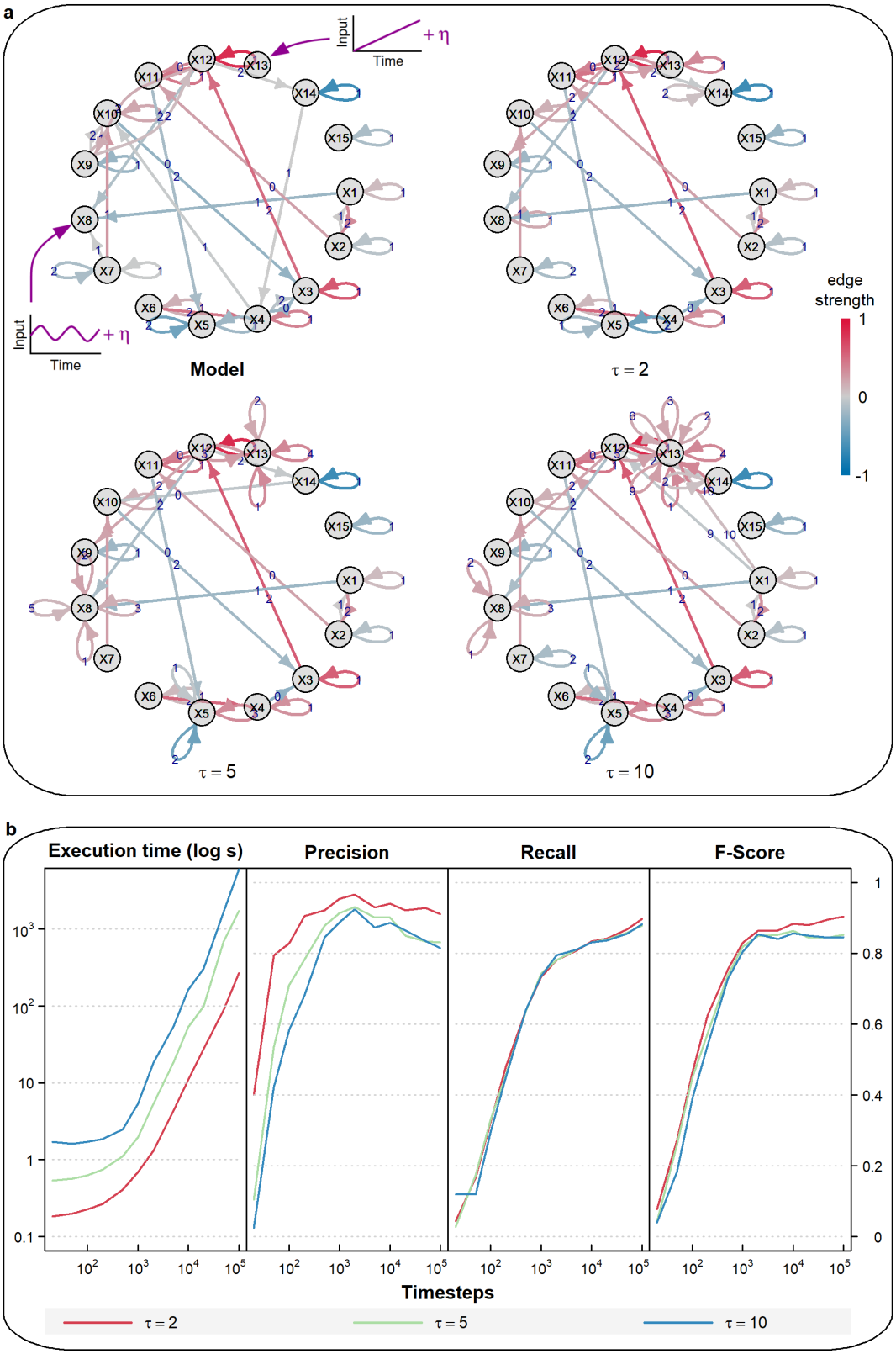
This work was supported by ITMO Cancer (grant No 20CM106) and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 847718. LD acknowledges support from AMX PhD fellowship, VC from ARC foundation and NL from CNRS-Imperial College joint PhD programme.



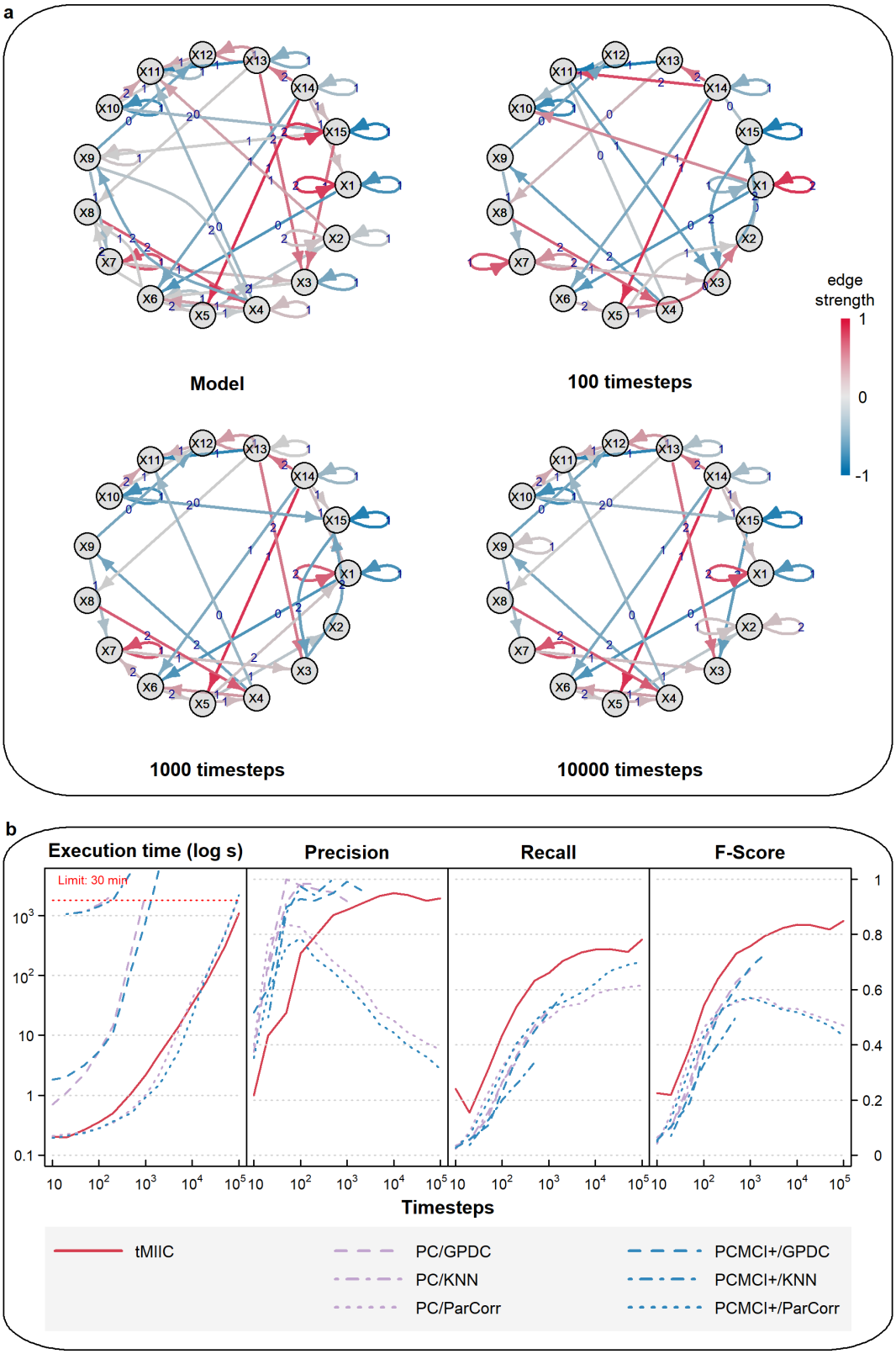
Extended Data Fig. 1: **Benchmark assessment of CausalXtract's causal discovery module (tMIIC) using generated time series datasets.** **a**, Example of a 15 node causal network to generate benchmark time series datasets based on linear combinations of contributions, Supplementary Table 1. Examples of temporal causal networks reconstructed by tMIIC based on 100, 1,000 or 10,000 simulated time steps. **b**, Running times and scores (Precision, Recall, Fscore) averaged over 10 datasets and compared to PC and PCMCI+ methods using different kernels (GPDC, KNN, ParCorr); tMIIC is at par with PC and PCMCI+ scores using GPDC and KNN kernels but runs orders of magnitude faster. Only ParCorr kernel matches tMIIC running speed but with significantly lower scores at large sample size, see Methods.



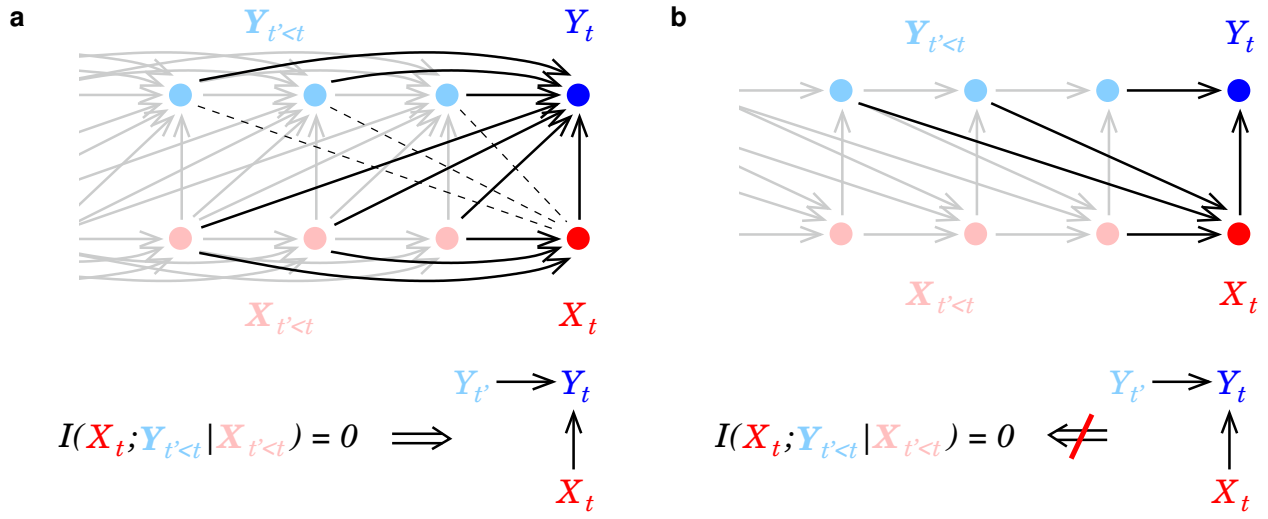
Extended Data Fig. 2: **CausalXtract insensitivity to an overestimated maximum lag τ. a**, Example of a temporal causal network model with a maximum lag $\tau = 2$. Corresponding temporal causal networks inferred by CausalXtract's causal discovery module (tMIIC), from 1,000 time step stationary time series (Supplementary Table 1), while assuming different maximum lags $\tau = 2, 5$ or 10. **b**, Running times and scores (Precision, Recall, Fscore) of tMIIC temporal causal network reconstructions for $\tau = 2, 5$ or 10, averaged over ten stationary time series of 10 to 10^5 time steps. Overestimating the maximum lag τ has little impact on the reconstructed networks, as long as the time series are stationary, as demonstrated in Extended Data Fig. 3.



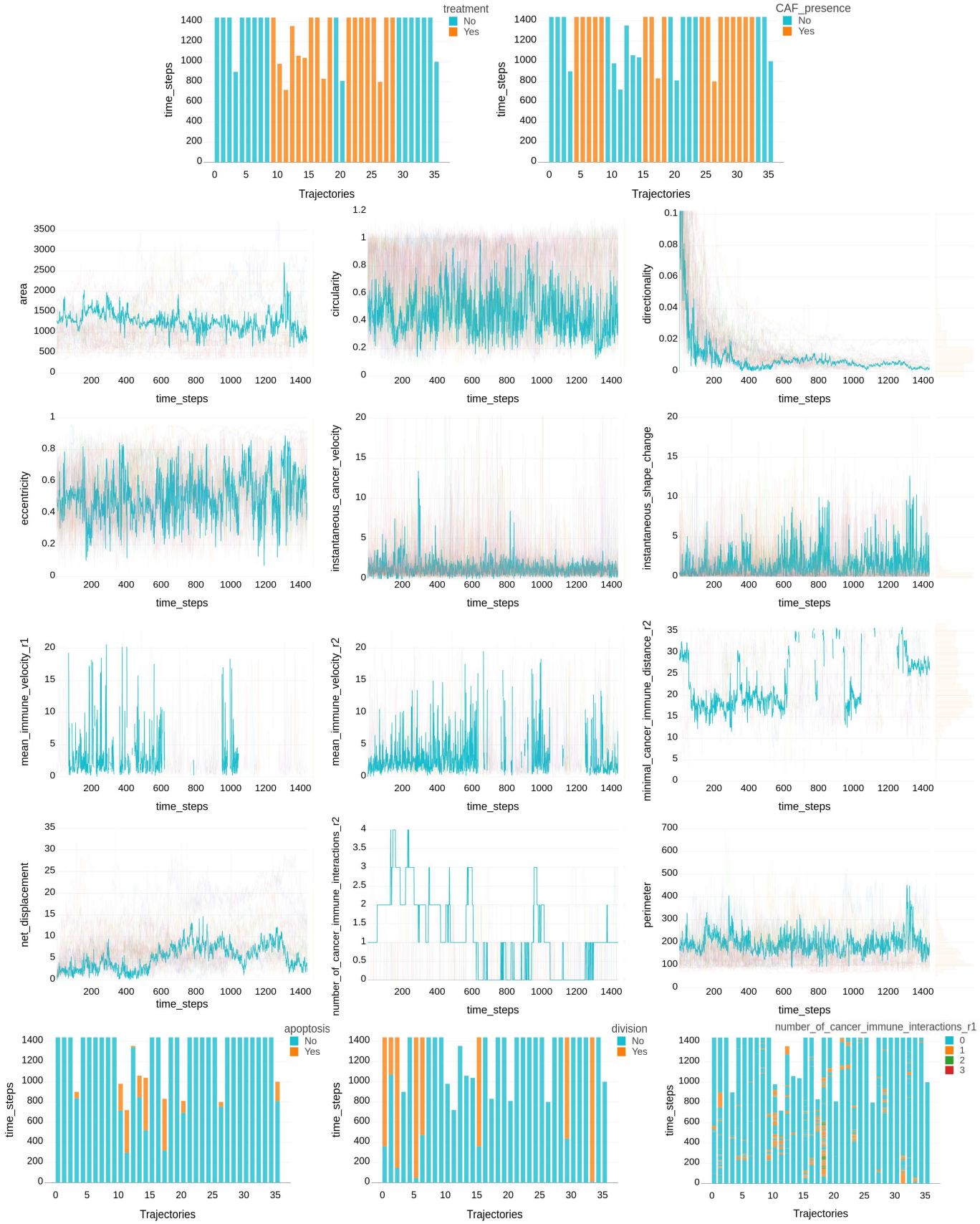
Extended Data Fig. 3: **CausalXtract sensitivity to non-stationary variables. a**, Example of a temporal causal network model ($\tau = 2$) with a low frequency periodic input ($T = 100$) applied to X8 and a time-linear trend applied to X13. Corresponding temporal causal networks inferred by tMIIC from 1,000 time step time series (Supplementary Table 1) including non-stationary inputs to X8 and X13. Increasing the maximum lag from $\tau = 2$ to $\tau = 5$ or 10 leads to the appearence of multiple self-loops, which result from the non-stationary dynamics of X8 and X13, whilst the rest of the network remains largely unaffected. **b**, Running times and scores (Precision, Recall, Fscore ignoring X8 and X13 self-loops) of tMIIC causal network reconstructions for $\tau = 2, 5$ or 10, averaged over ten time series of 10 to 10^5 time steps.



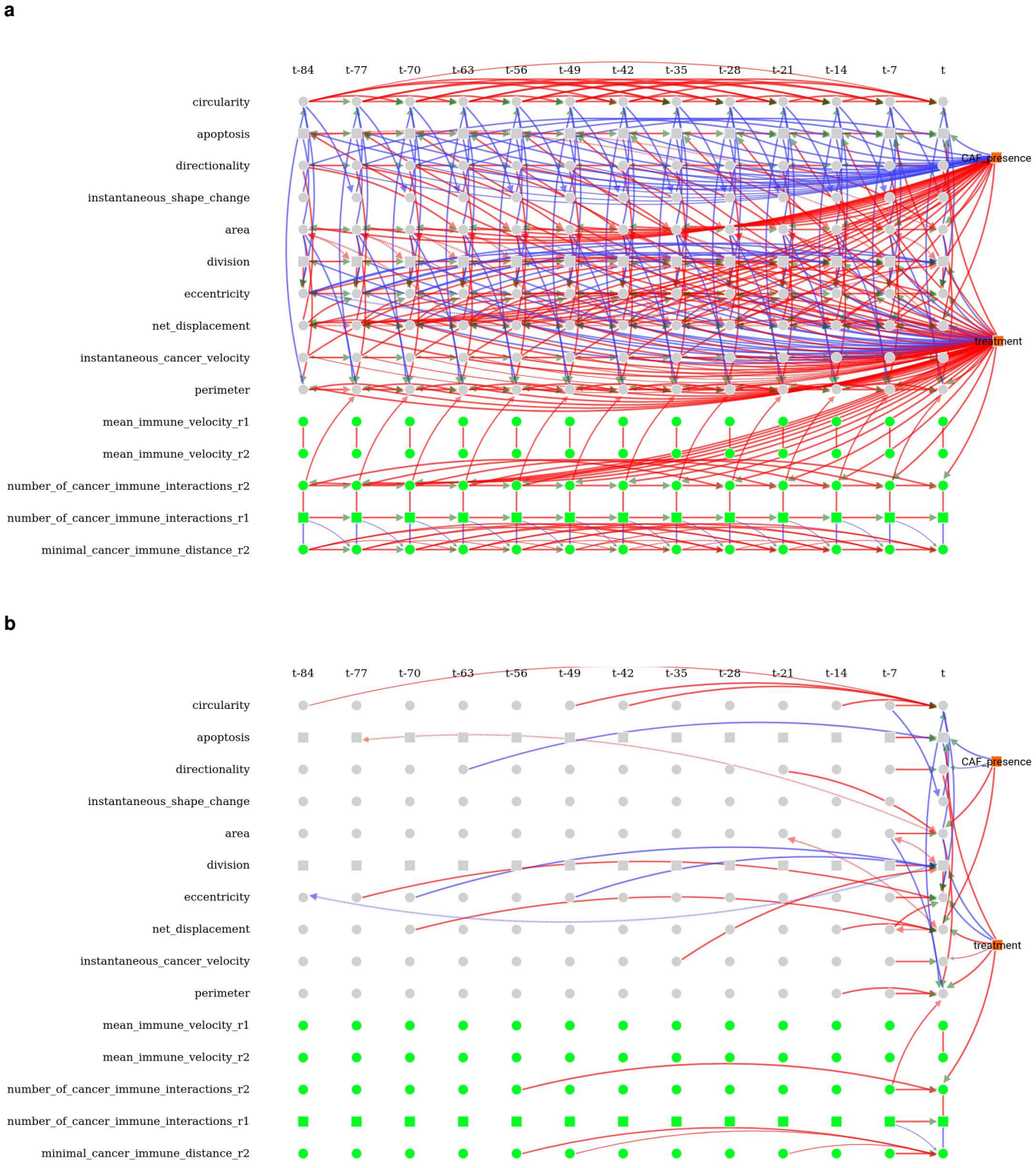
Extended Data Fig. 4: **Benchmark assessment of CausalXtract's causal discovery module (tMIIC) using more complex time series datasets.** **a**, Example of a 15 node causal network to generate more complex benchmark time series datasets based on non-linear combinations of contributions, Supplementary Table 2. Examples of temporal causal networks reconstructed by tMIIC based on 100, 1,000 or 10,000 simulated time steps. **b**, Running times and scores (Precision, Recall, Fscore) averaged over 10 datasets and compared to PC and PCMCI+ methods using different kernels (GPDC, KNN, ParCorr); tMIIC outperforms both PC and PCMCI+, in terms of Recall and Fscores, while running orders of magnitude faster, except for the ParCorr kernel, which leads, however, to significantly lower scores at large sample size.



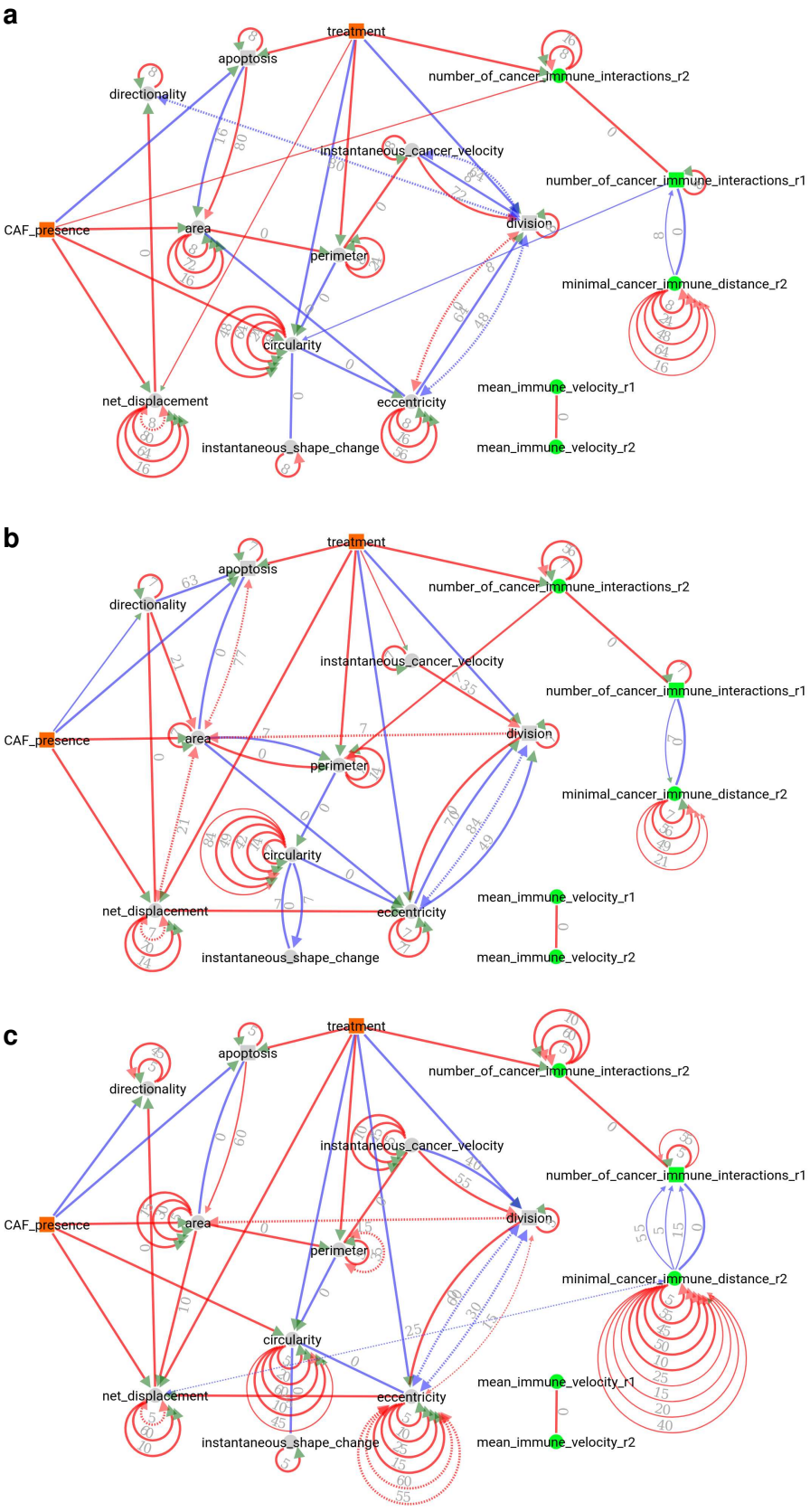
Extended Data Fig. 5: **Time-unfolded causal network framework and relation to Granger-Schreiber temporal causality.** **a**, A vanishing Transfer Entropy, *i.e.* $T_{Y\to X} = I(X_t; \boldsymbol{Y}_{t'<t}|\boldsymbol{X}_{t'<t}) = 0$, implies *i)* the absence of (dashed) edge between X_t and any $Y_{t'}$, with $t' < t$, and *ii)* if X_t is adjacent to Y_t, the presence of temporal (2-variable + time) v-structures, $Y_{t'} \to Y_t \leftarrow X_t$, for all $Y_{t'}$ adjacent to Y_t, with $t' < t$ (Methods, Theorem 1). These results can be readily extended to include the presence of other observed variables, $\boldsymbol{V}_{t'\leqslant t}$, by redefining Transfer Entropy as, $T_{Y\to X} = I(X_t; \boldsymbol{Y}_{t'<t}|\boldsymbol{X}_{t'<t}, \boldsymbol{V}_{t'\leqslant t})$, which discards contributions from indirect paths through other observed variables, $\boldsymbol{V}_{t'\leqslant t}$. **b**, By contrast, the presence of a temporal (2-variable + time) v-structure, $Y_{t'} \to Y_t \leftarrow X_t$ does not imply a vanishing Transfer Entropy, as long as there remains an edge between any $Y_{t''<t}$ and X_t. It implies that Granger-Schreiber temporal causality is in fact too restrictive and may overlook actual causal effects, which can be uncovered by graph-based causal discovery methods like CausalXtract's causal discovery module (tMIIC). Hence, CausalXtract's time-unfolded network framework, combining graph-based and information-based approaches, sheds light on the common foundations of the seemingly unrelated graph-based causality and Granger-Schreiber temporal causality, while clarifying their actual differences and limitations.



Extended Data Fig. 6: **Time series of cellular features extracted from the tumor ecosystems.** Example of time series of cellular features extracted by CausalXtract's feature extraction module (CellHunter+) from the tumor ecosystems analyzed in this study, Fig. 1a. It includes two experimental control parameters (*i.e.* treatment and CAF presence) and 15 cellular features extracted every 2 minutes over a period of two days. Continuous features are highlighted for one trajectory (traj.18), while categorical features are shown for all trajectories.



Extended Data Fig. 7: **Time-unfolded causal network inferred by CausalXtract.** **a**, Time-unfolded causal network assuming stationary dynamics of cellular ecosystems implying translational time invariance of the inferred causal network. **b**, Only edges involving at least one contemporaneous variables (*i.e.* at time t) need to be tested for conditional independence by tMIIC and the remaining edges are then duplicated at all previous time steps before assigning orientations when time-lagged latent variables are taken into account, Fig. 1c. Variables retaining multiple self-loops with different time-delays correspond to non-stationary variables in Extended Data Fig. 6, in agreement with benchmarks from simulated data including non-stationary variables, Extended Data Fig. 3.



Extended Data Fig. 8: **Robustness of CausalXtract's temporal causal networks to variations in sampling rate.** Summary causal networks inferred by CausalXtract using different sampling rates ($\delta\tau$). **a**, $\delta\tau = 8$ ts and $\tau = 80$ ts, in time step units (1 ts = 2 min). **b**, $\delta\tau = 7$ ts, and $\tau = 84$ ts, as chosen automatically by CausalXtract based on the average relaxation time across the 15 monitored variables, $\tau_R = 40$ ts, which defines a maximum lag $\tau = 2\,\tau_R = 80$ ts. Given a total number of (time-lagged and -unlagged) nodes, chosen to be around 200 nodes for computational efficiency, it leads to 13 temporal layers ($\nu + 1 = 200/15 \simeq 13$) and a lag increment $\delta\tau = \tau/\nu \simeq 7$ ts. This summary causal network corresponds to Fig. 2a. **c**, $\delta\tau = 5$ ts and $\tau = 60$ ts, corresponding to $\tau = \nu \cdot \delta\tau$ with $\nu + 1 = 13$ temporal layers, as in (**b**).

Supplementary Table 1: 15 nodes model.

Nodes

$$
\begin{aligned}
X_t^1 &\leftarrow -0.47\ f_2(X_{t-1}^1) + 0.29\ f_3(X_{t-1}^2) \times \eta_1 \\
X_t^2 &\leftarrow 0.49\ f_2(X_{t-1}^2) + 0.4\ f_1(X_{t-2}^1) + \eta_2 \\
X_t^3 &\leftarrow 0.56\ f_1(X_{t-1}^3) + 0.44\ f_4(X_{t-2}^4) - 0.26\ f_2(X_{t-2}^{10}) + 0.56\ f_2(X_t^4) + \eta_3 \\
X_t^4 &\leftarrow 0.24\ f_3(X_{t-1}^4) - 0.24\ f_2(X_{t-2}^6) - 0.12\ f_4(X_{t-1}^{14}) \times \eta_4 \\
X_t^5 &\leftarrow -0.39\ f_3(X_{t-1}^5) - 0.42\ f_3(X_{t-2}^5) - 0.39\ f_3(X_t^{11}) + \eta_5 \\
X_t^6 &\leftarrow -0.32\ f_2(X_{t-1}^6) + \eta_6 \\
X_t^7 &\leftarrow -0.17\ f_4(X_{t-1}^7) - 0.17\ f_1(X_{t-2}^7) + \eta_7 \\
X_t^8 &\leftarrow 0.39\ f_4(X_{t-1}^8) - 0.46\ f_4(X_{t-1}^7) - 0.39\ f_3(X_{t-1}^1) - 0.4\ f_3(X_{t-2}^{12}) + \eta_8 \\
X_t^9 &\leftarrow -0.34\ f_1(X_{t-1}^9) + 0.43\ f_3(X_{t-2}^{12}) + \eta_9 \\
X_t^{10} &\leftarrow 0.2\ f_1(X_{t-1}^{10}) + 0.18\ f_4(X_{t-2}^9) + 0.17\ f_1(X_{t-1}^9) + 0.48\ f_3(X_{t-1}^7) - 0.26\ f_4(X_{t-1}^4) + \eta_{10} \\
X_t^{11} &\leftarrow 0.41\ f_2(X_{t-1}^{11}) + 0.54\ f_3(X_t^2) - 0.55\ f_2(X_t^{12}) + \eta_{11} \\
X_t^{12} &\leftarrow -0.45\ f_2(X_{t-1}^{12}) - 0.43\ f_4(X_{t-2}^3) - 0.17\ f_4(X_{t-2}^9) \times \eta_{12} \\
X_t^{13} &\leftarrow 0.45\ f_3(X_{t-1}^{13}) + \eta_{13} \\
X_t^{14} &\leftarrow 0.28\ f_2(X_{t-1}^{14}) + 0.37\ f_1(X_{t-2}^{12}) \times \eta_{14} \\
X_t^{15} &\leftarrow 0.52\ f_3(X_{t-1}^{15}) + \eta_{15}
\end{aligned}
$$

Functions

$$
\begin{aligned}
f_1(x) &= x \\
f_2(x) &= x\,(1 - 4\,e^{-\frac{x^2}{2}}) \\
f_3(x) &= x\,(1 - 4\,x^3\,e^{-\frac{x^2}{2}}) \\
f_4(x) &= \cos(x)
\end{aligned}
$$

Noises

The η are white noises generated for each node or contribution using a normal distribution:

$$\eta \sim \mathcal{N}(0, 1)$$

Supplementary Table 2: 15 nodes model with combinations.

Nodes

$$
\begin{aligned}
X_t^1 &\leftarrow \eta - 0.7\, f_6(u(\eta + X_{t-1}^1)) - 0.87\, f_5(u(\eta + (X_{t-1}^{14} \times X_{t-2}^1))) \\
X_t^2 &\leftarrow \eta + 0.65\, f_1(u(\eta + X_{t-1}^2)) - 0.63\, f_3(u(\eta + X_{t-2}^2)) + 0.79\, f_3(u(\eta + X_{t-1}^5)) \\
X_t^3 &\leftarrow \eta - 0.76\, f_5(u(\eta + X_{t-1}^3)) - 0.59\, f_6(u(\eta + X_{t-1}^7)) - 0.85\, f_2(u(\eta + X_{t-1}^{15})) \\
&\quad -0.89\, f_5(u(\eta + (X_{t-2}^{13} \times X_{t-1}^7))) \\
X_t^4 &\leftarrow \eta - 0.7\, f_6(u(\eta + X_{t-1}^5)) - 0.86\, f_2(u(\eta + X_{t-2}^8)) + 0.53\, f_1(u(\eta + (X_{t-1}^4 \times X_{t-2}^9))) \\
X_t^5 &\leftarrow \eta + 0.54\, f_2(u(\eta + (X_{t-1}^{14} \times X_{t-2}^6))) \\
X_t^6 &\leftarrow \eta - 0.85\, f_2(u(\eta + X_{t-1}^6)) - 0.79\, f_3(u(\eta + X_{t-2}^3)) + 0.59\, f_1(u(\eta + X_{t-1}^4)) \\
&\quad +0.75\, f_3(u(\eta + X_t^1)) + 0.57\, f_2(u(\eta + X_{t-1}^{14})) \\
X_t^7 &\leftarrow \eta + 0.74\, f_1(u(\eta + X_{t-1}^7)) + 0.54\, f_6(u(\eta + X_{t-1}^9)) - 0.53\, f_2(u(\eta + (X_{t-1}^9 \times X_{t-1}^7))) \\
X_t^8 &\leftarrow \eta \times (-0.63\, f_1(u(\eta + X_{t-1}^6)) + 0.81\, f_5(u(\eta + X_t^{13})) + 0.53\, f_6(u(\eta + (X_{t-2}^6 \times X_{t-1}^6))) \\
&\quad -0.69\, f_6(u(\eta + (X_t^{13} \times X_{t-1}^6)))) \\
X_t^9 &\leftarrow \eta + 0.79\, f_3(u(\eta + X_{t-2}^4)) + 0.69\, f_6(u(\eta + (X_{t-1}^9 \times X_{t-1}^{15}))) \\
X_t^{10} &\leftarrow \eta + 0.54\, f_6(u(\eta + X_{t-1}^{10})) \\
X_t^{11} &\leftarrow \eta + 0.83\, f_6(u(\eta + X_{t-1}^{11})) - 0.76\, f_4(u(\eta + X_{t-1}^{13})) - 0.73\, f_3(u(\eta + X_{t-1}^2)) \\
&\quad +0.74\, f_2(u(\eta + X_t^4)) - 0.87\, f_2(u(\eta + X_{t-2}^{10})) + 0.72\, f_4(u(\eta + X_{t-1}^{12})) \\
&\quad -0.73\, f_1(u(\eta + (X_{t-2}^{10} \times X_{t-1}^{13}))) \\
X_t^{12} &\leftarrow \eta + 0.7\, f_3(u(\eta + X_{t-1}^{10})) - 0.55\, f_5(u(\eta + X_t^9)) - 0.54\, f_5(u(\eta + (X_{t-1}^{12} \times X_{t-1}^{10}))) \\
X_t^{13} &\leftarrow \eta - 0.62\, f_3(u(\eta + X_{t-2}^{14})) - 0.61\, f_1(u(\eta + (X_{t-1}^{13} \times X_{t-2}^{14}))) \\
X_t^{14} &\leftarrow \eta - 0.78\, f_6(u(\eta + X_{t-1}^{14})) \\
X_t^{15} &\leftarrow \eta - 0.68\, f_4(u(\eta + X_{t-1}^{15})) + 0.85\, f_4(u(\eta + X_{t-2}^{15})) - 0.6\, f_5(u(\eta + X_{t-2}^{10})) \\
&\quad +0.68\, f_6(u(\eta + X_{t-1}^{14})) + 0.81\, f_4(u(\eta + (X_{t-1}^{14} \times X_{t-2}^{10})))
\end{aligned}
$$

Functions

$$
\begin{aligned}
u(x) &= \max(-1, \min(1, x)) \\
f_1(x) &= x \\
f_2(x) &= x\,(1 - 4\,e^{-\frac{x^2}{2}})/1.52387 \\
f_3(x) &= 4\,x^2 \\
f_4(x) &= 8\,x^3 \\
f_5(x) &= 16\,x^4 \\
f_6(x) &= \cos(\pi x)
\end{aligned}
$$

Noises

The η are white noises generated for each node or contribution using a normal distribution:

$$\eta \sim \mathcal{N}(0, 0.1)$$

Publications en co-auteur

Learning interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients

Marcel da Câmara Ribeiro-Dantas[a,1], **Honghao Li**[a,1], **Vincent Cabeli**[a,1], **Louise Dupuis**[a,1], **Franck Simon**[a], **Liza Hettal**[a], **Anne-Sophie Hamy**[b,c,d], **and Hervé Isambert**[2,a]

[a]CNRS UMR168, Institut Curie, Université PSL, Sorbonne Université, Paris, France; [b]INSERM U932, Institut Curie, Paris, France; [c]Department of Medical Oncology, Institut Curie, Saint-Cloud, France; [d]Department of Surgery, Institut Curie, Université Paris, Paris, France

This manuscript was compiled on March 14, 2023

Discovering causal effects is at the core of scientific investigation but remains challenging when only observational data is available. In practice, causal networks are difficult to learn and interpret, and limited to relatively small datasets. We report a more reliable and scalable causal discovery method (iMIIC), based on a general mutual information supremum principle, which greatly improves the precision of inferred causal relations while distinguishing genuine causes from putative and latent causal effects. We showcase iMIIC on synthetic and real-life healthcare data from 396,179 breast cancer patients from the US Surveillance, Epidemiology, and End Results program. More than 90% of predicted causal effects appear correct, while the remaining unexpected direct and indirect causal effects can be interpreted in terms of diagnostic procedures, therapeutic timing, patient preference or socio-economic disparity. iMIIC's unique capabilities open up new avenues to discover reliable and interpretable causal networks across a range of research fields.

Causal discovery | Interpretable causal networks | Healthcare data | Breast cancer

Nationwide medical records contain massive amounts of real-life data on human health, including some personal, familial and socio-economic information, which frequently affect not only health conditions, but also timing of diagnosis, medical treatments and, ultimately, the survival of patients. Besides, such non-medical determinants of human health are usually controlled for in clinical trials, which select specific groups of patients through restrictive enrolment criteria. Yet, the wealth of information contained in real-life medical records remains largely under-exploited due to the lack of unsupervised methods and tools to analyze them without preconceived hypotheses. This highlights the need to develop new machine learning strategies to analyze healthcare data, in order to uncover unsuspected associations and possible cause-effect relations between all available information recorded in the medical history of patients, Fig. 1a.

Learning cause-effect relations from purely observational data has long been known to be, in principle, possible thanks to seminal works on causal discovery methods (1, 2). In essence, causal discovery infers cause-effect relations from specific correlation patterns involving at least three variables, which goes beyond the popular notion that pairwise correlation does not imply causation. However, while observational data account for the vast majority of available datasets across a wide range of domains, uncovering cause-effect relations still remains notoriously challenging in absence of systematic intervention, which might be impractical, too costly or unethical, when it concerns human health.

While causal discovery is tightly linked to methods designed to learn graphical models (1–4), most structure learning methods are not actually designed to uncover cause-effect relations. In particular, maximum likelihood approaches, such as Search-and-Score (5) or Graphical Lasso (6) methods, are restricted to specific model classes, assuming either fully directed graphs or fully undirected graphs, and cannot therefore learn the causal or non-causal nature of graph edges. By contrast, constraint-based causal discovery methods assume broader classes of graphs and can learn the orientation of certain edges solely based on observational data (1, 2), Fig. 1b. To this end, they first learn structural constraints, in the form of conditional independence relations, which provide indirect and somewhat cryptic information about possible causal relationships between observed as well as unobserved variables, as outlined in Box 1. Yet, despite being theoretically sound given unlimited amount of data (7), constraint-based methods remain unreliable and difficult to interpret on the relatively small datasets, they can handle in practice.

We report here the advanced causal discovery method, iMIIC (interpretable MIIC), that can learn more reliable and interpretable causal graphical models, as well as, handle much larger datasets (*e.g.* including a few hundred thousand samples). The novel iMIIC method greatly expands the causal discovery performance of the recent structure learning method, MIIC (Multivariate Information-based Inductive Causation), combining constraint-based and information-theoretic frameworks (8–10). iMIIC's performance relies on three main con-

Author Summary

Uncovering causal relations is notoriously challenging when only observational data is available, as with the vast majority of datasets for which systematic perturbation experiments are not feasible. We report a reliable causal discovery method, iMIIC, which can uncover genuine causal relations from multi-variable correlations in biological, biomedical or other complex observational data. Based on information theory principles, iMIIC can analyze very large multimodal datasets integrated from different sources or experimental techniques. We demonstrate iMIIC causal discovery performance on healthcare data from 396,179 breast cancer patients. iMIIC's unique graphical output provides a global view of the direct and indirect relations between variables, including some unexpected cause-effect findings, which can be interpreted in terms of clinical practice, patient choices or socio-economic background.

M.C.R-D., H.L., V.C., F.S. and H.I. designed and implemented the machine learning tools; M.C.R-D., H.L., L.D. contributed to data analysis; M.C.R-D., H.L., L.H., A-S.H. and H.I. contributed to data interpretation. H.I., L.D., M.C.R-D., H.L. wrote the manuscript.

The authors declare no competing interest.

[1]M.C.R-D., H.L., V.C. and L.D. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: herve.isambertcurie.fr

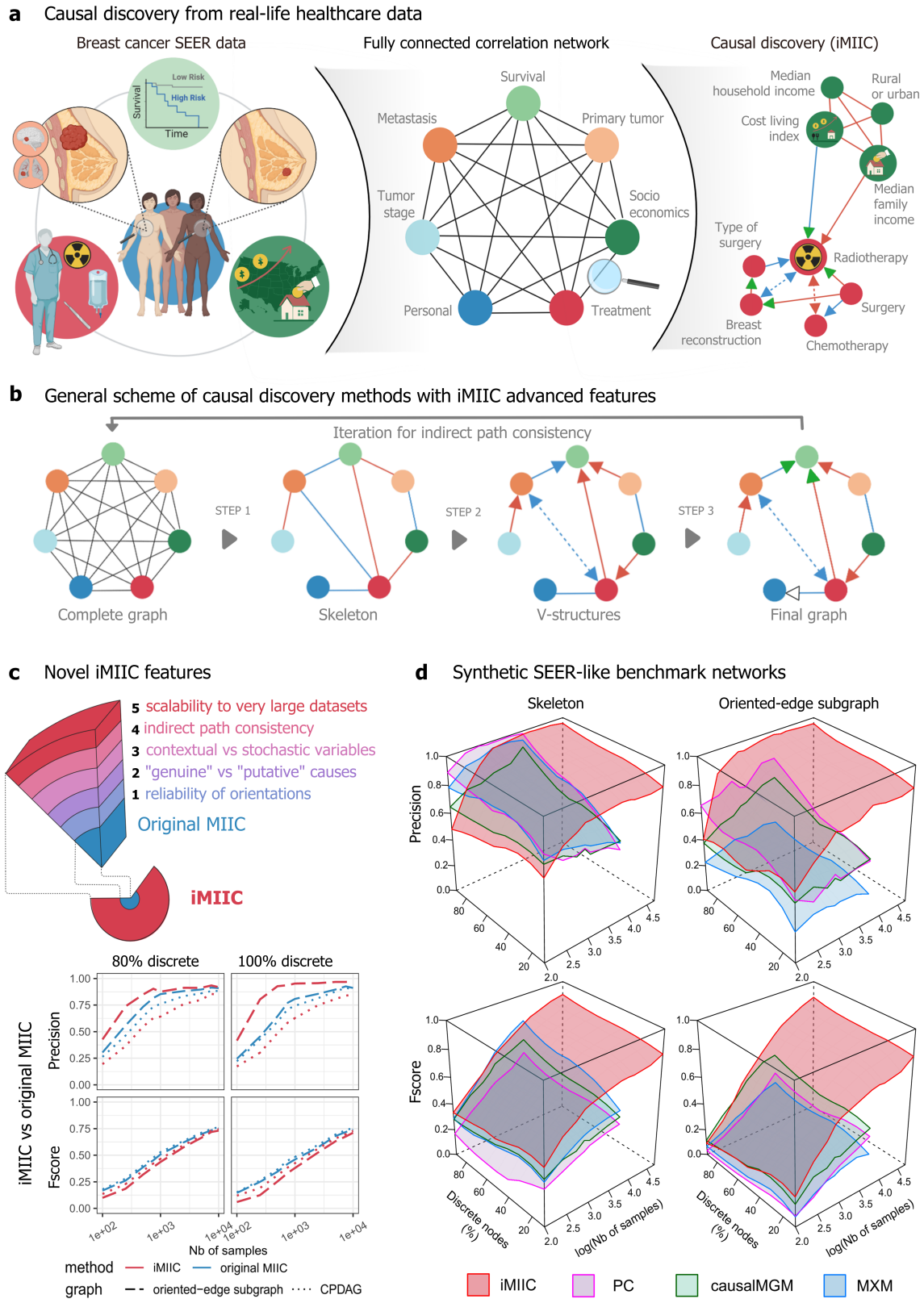


Fig. 1. Causal discovery from real-life healthcare data using constraint-based methods. **(a)** SEER database includes 407,791 medical records of breast cancer patients diagnosed between 2010 and 2016. Causal discovery aims at uncovering cause-effect relations across such globally correlated datasets. **(b)** General scheme of constraint-based methods (including iMIIC's novel advanced features, see main text and *SI Appendix*): Step 1, removal of dispensable edges (guaranteeing indirect path consistency); Step 2, 'v-structure' orientation (with reliable orientations and latent common causes shown as bidirected edges); Step 3, propagation of orientation shown with white arrowhead (and distinction between 'putative' and 'genuine' causes, green arrowheads). **(c)** Novel iMIIC advanced features and benchmark comparison with original MIIC. **(d)** Synthetic SEER-like benchmark networks with different proportions of discrete variables, see text, Materials and Methods and Figs. S4-S6. Created with BioRender.com

ceptual advances and associated methodological developments. First, iMIIC quantitatively improves the reliability of inferred orientations, based on a general information-theoretic principle. It results in only a few percents of false positive orientations on challenging benchmarks adapted from real-life healthcare data. Second, iMIIC is uniquely able to distinguish "genuine" causes from "putative" and "latent" causal effects. This is an essential distinction to disambiguate the causal interpretation of oriented edges in inferred networks, as outlined on an intuitive example in Box 1. Third, iMIIC quantifies indirect effects, while ensuring their consistency with the global network structure. This is important to interpret indirect contributions in term of indirect paths through the corresponding contributor nodes in the inferred network, which is generally not possible with other causal discovery methods. In addition, iMIIC distinguishes contextual from stochastic variables, which allows the inclusion of externally controlled variables in causal networks, and, finally, iMIIC enables scalability to very large datasets. These unique capabilities open up new avenues to discover reliable and interpretable causal networks across a range of research fields. We demonstrate iMIIC's causal discovery performance on synthetic and real-life healthcare data originating from more than 400,000 medical records of breast cancer patients from the Surveillance, Epidemiology, and End Results (SEER) program (11), Fig. S1.

Results

Overview and limitations of causal discovery methods. Constraint-based causal discovery methods proceed through successive steps, Fig. 1b (*SI Appendix*, section 1). The first step consists in removing, iteratively, all dispensable edges from an initial fully connected network, whenever two variables are independent or conditionally independent given a so-called "separating set" of conditioning variables. The second step then consists in orienting some of the edges of the undirected graph (named skeleton) to form so-called "v-structures", $X \to Z \leftarrow Y$, which are the signature of causality in observational data, Box 1. Finally, the third step aims at propagating the orientations of v-structures to downstream edges, Fig. 1b. However, traditional constraint-based methods lack robustness on finite datasets, as their long series of uncertain decisions lead to an accumulation of errors, which limit the reliability of the final networks. In particular, spurious conditional independences, stemming from coincidental combinations of conditioning variables, lead to many false negative edges and, ultimately, limit the accuracy of inferred orientations.

The recent causal discovery method, MIIC (8, 10), learns more robust causal graphical models by first collecting iteratively significant information contributors before assessing conditional independences (*SI Appendix*, section 2). In practice, MIIC's strategy limits spurious conditional independences and significantly improves the sensitivity or recall (*i.e.*, the fraction of correctly recovered edges) compared to traditional constraint-based methods, Figs. S2 and S3. Yet, original MIIC as well as all other existing causal discovery methods still present a number of limitations. In particular, (*i*) they present a lower reliability in predicting edge orientation than edge retention, (*ii*) they uncover "putative" rather than "genuine" causal relations (Box 1), (*iii*) they do not guarantee indirect path consistency with the global network structure, (*iv*) they do not distinguish contextual from stochastic variables, and (*v*) they have a limited scalability. The novel iMIIC method effectively overcomes all these limitations and greatly enhances the reliability, interpretability and scalability of causal discovery on large scale synthetic as well as real-life observational data.

iMIIC improves the reliability of inferred orientations. While the original MIIC significantly outperforms traditional constraint-based methods in inferring reliable orientations, a substantial loss in precision usually remains between MIIC skeleton and oriented graph predictions, Fig. S3. This is due to orientation errors originating mainly from inconsistent v-structures, $X \to Z \leftarrow Y$, whose middle node Z could also be included in the separating set of the unconnected pair $\{X, Y\}$, in contradiction with the head-to-head meeting of the v-structure. Inconsistent v-structures are particularly common for datasets including discrete variables with (too) many levels. To prevent such inconsistent orientations, iMIIC implements more conservative orientation rules, based on a general mutual information supremum principle (15, 16), regularized for finite datasets (*SI Appendix*, section 3). This principle implies to aggregate the levels of categorical or continuous variables alike, when assessing (conditional) independence. As a result, Theorem 1 (*SI Appendix*, section 3) requires to rectify all (conditional) mutual information between independent variables. Combined with more scalable computations of multivariate information and orientation scores (*SI Appendix*, section 4), this information-theoretic principle greatly enhances the reliability of predicted orientations with only a small sensitivity loss compared to MIIC original orientation rules, Fig. 1c. In particular, iMIIC's orientation precision exceeds 90% on challenging benchmarks adapted from real-life heterogeneous data, outlined below, when other causal discovery methods typically level off below 50-60% orientation precision at large sample size, Fig. 1d (oriented-edge subgraph precision plot) and Figs. S4-S6 (dashed lines in precision plots).

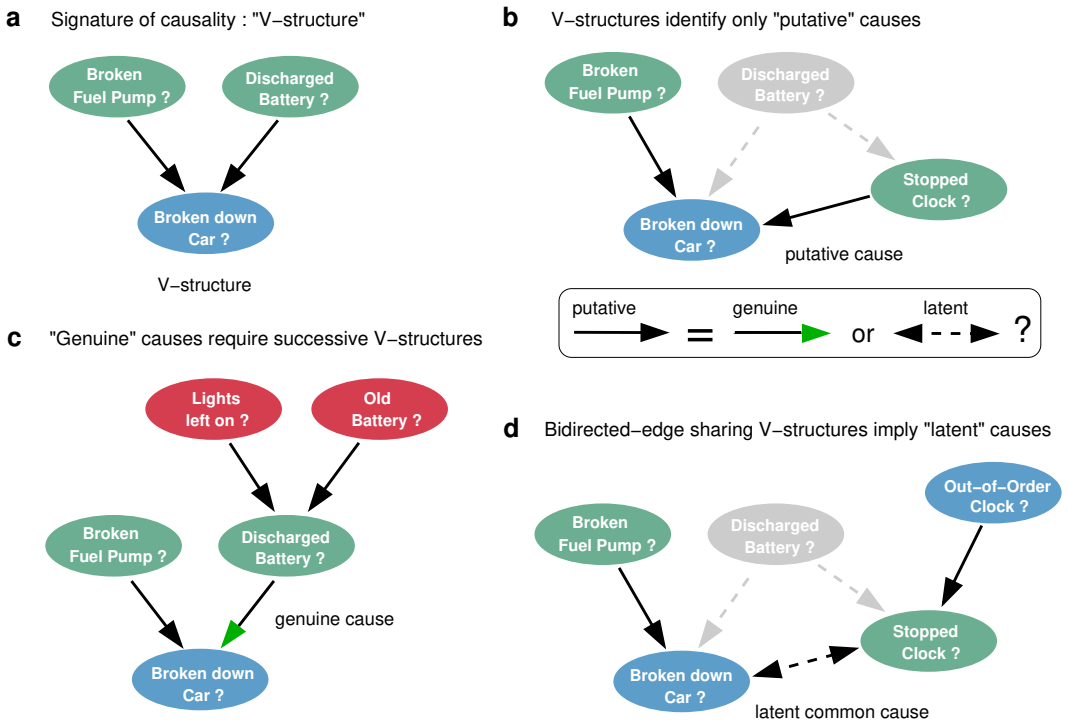
iMIIC distinguishes "genuine" from "putative" causal edges. Traditional constraint-based methods and indeed the original MIIC method merely discover "putative" causal relations, as v-structure orientations are *a priori* compatible with both genuine cause-effect relations and the effects of unobserved common causes, as outlined on an intuitive example in Box 1. By contrast, iMIIC distinguishes "genuine" from "putative" causal edges by ruling out the effect of an unobserved common cause (or unmeasured confounder) for each predicted genuine causal edge. This unique feature of iMIIC is achieved by assessing separate probabilities of arrow head and tail for all oriented edges (*SI Appendix*, section 4). Genuine causal edges (represented with a green arrow head) are then predicted if both arrow head and tail probabilities are statistically significant, while causal edges remain "putative" if their tail probability is not statistically significant or cannot be determined from purely observational data. Likewise, bidirected edges, interpreted as the effect of unobserved common causes, correspond to two significant head probabilities, while all other cases are graphically represented as undirected edges (*SI Appendix*, section 5).

iMIIC allows for contextual variables in causal networks. The separate probabilistic framework of arrow head *versus* tail orientations implemented in iMIIC also allows to include prior knowledge about certain head or tail orientations. For instance,

Box 1. Causal discovery principles from observational data: distinguishing "genuine" causes from "putative" and "latent" causal effects.

We outline here the principles to uncover cause-effect relations in a purely observational dataset and distinguish "genuine" causes from "putative" and "latent" causes. The rationale is illustrated on the causally intuitive toy example of an imaginary dataset of old cars. (**a**) The signature of causality in such observational datasets corresponds to 3-variable "v-structure" subgraphs involving two *independent* and thus *unconnected* possible causes, "Broken fuel pump?" and "Discharged battery?", and a resulting effect, "Broken down car?". The converging orientations of this v-structure towards its middle variable, "Broken down car?", stem from the fact that these two edges cannot be undirected, nor can they point towards either "Broken fuel pump?" or "Discharged battery?", as these alternative graphical models would imply correlations contradicting the independence between "Broken fuel pump?" and "Discharged battery?". Alternatively, causal relations can sometimes be uncovered between two variables only, under the specific assumption of continuous additive noise models (12). However, in the general case, causal discovery requires at least three and often more variables, as the independence between possible causes in a v-structure is frequently conditional on other variable(s), not considered here, defining a separating set, see *SI Appendix*, section 1. Conversely, conditioning on the tip of a v-structure, here "Broken down car?", induces spurious associations between its independent possible causes (1, 2). Likewise, selecting a dataset with specific values for this tip variable results in spurious associations due to selection bias in the dataset (13, 14), such as some apparent anti-correlation between different possible causes, "Broken fuel pump?" and "Discharged battery?", if only "Broken down car? = yes" are selected. (**b**) However, v-structures remain in fact causally ambiguous (2) as they only identify "putative" causes, which can either be "genuine" causes, displayed with a green arrowhead, or suggest the presence of unmeasured confounders, *i.e.* latent common causes unobserved in the dataset and represented with a bidirected edge. For instance, the variable "Clock stopped?", frequently used as a proxy for "Discharged battery?", also forms a similar v-structure with "Broken fuel pump?"; yet, it is well known that "Clock stopped?" cannot be a genuine cause of "Broken down car?", as tampering with a car's clock cannot actually cause a car to break down. (**c**) In absence of background knowledge and direct intervention on variables, showing that "Discharged battery?" is indeed a genuine cause of "Broken down car?" requires to exclude the possibility of an unobserved common cause (*i.e.* an unmeasured confounder) between "Discharged battery" and "Broken down car?". To this end, one needs to find another v-structure upstream of "Discharged battery?" (*e.g.* "Lights left on?"→"Discharged battery?"←"Old battery?") or to have prior knowledge about an upstream (putative) cause and to show that the effect of at least one upstream variable on the downstream variable "Broken down car?" is entirely *indirect* and mediated (at least in part) by the intermediary variable "Discharged battery?". This requires to find a conditional independence between an upstream variable and "Broken down car?" conditioned on a separating set, which includes the intermediary variable "Discharged battery". (**d**) Conversely, ruling out a putative cause as genuine cause requires to show that the relation actually originates from an unobserved common cause by finding a fourth variable (*e.g.* "Out-of-order clock?") defining another v-structure, inducing a bidirected edge between "Broken down car?" and "Clock stopped?" with the v-structure in (b).

The advanced iMIIC method distinguishes genuine from putative causal edges, as well as, undirected and bidirected edges, by assessing separate head or tail orientation probabilities at each edge extremity (see Results and *SI Appendix*, sections 4 and 5). Hence, iMIIC can discover four types of edges with different causal interpretations. iMIIC provides also estimates of direct and indirect information contributions between any pair of variables, *SI Appendix*, section 6. However, like other causal discovery methods, iMIIC does not *quantify* causal effects, which requires additional assumptions (identifiability), not generally testable in observational studies (2). In particular, the causal effects of a putative cause are nonidentifiable, implying that the results of intervention on a putative cause cannot be quantified from observational data alone.



including a few contextual variables in graphical models can help specify a control parameter or experimental conditions or characterize the personal profile of patients (*e.g.* sex, year of birth), depending on the nature of the dataset. Unlike most other variables of the dataset, such contextual variables are not stochastically varying and should have, by assumption, all their edges without incoming arrow head, *i.e.*, $p_{\text{tail}} = 1$. This expresses our prior knowledge that contextual variables cannot be the consequence of other observed or unobserved variables in the dataset.

iMIIC enforces indirect path consistency and quantifies their information contributions. The rationale behind the removal of dispensable edges in the first step of constraint-based causal discovery methods is that all statistical associations between disconnected variables should be graphically interpretable in terms of indirect paths in the final network. However, this is frequently not the case in practice (17). In particular, there is no guarantee that the separating sets identified during this iterative removal of edges remain consistent in terms of indirect paths in the final network. To this end, iMIIC adapts a novel

algorithmic scheme (17) to ensure that all separating sets identified to remove dispensable edges are consistent with the final inferred graph. It is achieved by repeating the constraint-based structure learning scheme, iteratively, while selecting only separating sets that are consistent with the skeleton or the partially oriented graph obtained at the previous iteration, as outlined in Fig. 1b. This indirect path consistency improves the interpretability of iMIIC inferred networks in terms of indirect effects, which are also quantified through indirect information contributions (*SI Appendix*, section 6).

iMIIC outperforms existing methods on synthetic benchmark datasets. The performance of iMIIC has been benchmarked against original MIIC as well as other state-of-the-art constraint-based methods on multiple SEER-like benchmark datasets with different proportions of discrete variables, see Materials and Methods. Fig. 1c demonstrates that iMIIC significantly improves the precision of orientations to the expense of a relatively small loss in orientation sensitivity and F-score for SEER-like benchmark datasets with large proportions of discrete variables. For instance, for $N = 500$ samples, orientation precision (resp. F-score) already reaches 93% (resp. 25%) with iMIIC *versus* 64% (resp. 35%) with original MIIC, for fully discrete SEER-like datasets, and exceeds 85% (resp. 32%) with iMIIC *versus* 73% (resp. 39%) with original MIIC, for 80% discrete variables as in the actual SEER dataset, Fig. 1c. In addition, iMIIC greatly outperforms the reliability and sensitivity of inferred orientations against other state-of-the-art constraint-based methods, Fig. 1d and Figs. S4-S6. In particular, iMIIC's orientation F-scores are about twice as high as PC algorithm's (18, 19) orientation F-scores, for all sample sizes and discrete variable proportions in these SEER-like datasets. For instance, for benchmarks with 80% discrete variables as in the actual SEER dataset, iMIIC already reaches 88% (resp. 44%) in precision (resp. F-score) for $N = 10^3$, against about 60% (18%) for conservative PC(19, 20), 50% (36%) for causalMGM(21) and 24% (18%) for MXM(22). For $N = 10^4$, iMIIC reaches 92% (73%) in precision (F-score), against about 75% (40%) for conservative PC, 62% (55%) for causalMGM and 30% (30%) for MXM. Finally, iMIIC reaches more than 90% for both orientation precision and F-score, for $N = 10^5$, which is beyond the sample size attainable by other methods. See Materials and Methods for comparisons with higher proportion of continuous variables.

Application to nationwide breast cancer medical records. We applied iMIIC on a large breast cancer dataset (11) from the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute, which collects data on cancer diagnoses, treatment and survival for ∼35% of the US population, Fig. 1a. Breast cancer (23) is the most common invasive cancer in women and is curable in only 70-80% of patients with large disparities in terms of tumor subtypes and stages at diagnostic, initial and subsequent treatments, as well as patient's age, ethnicity, genetic predisposition, lifestyle or socio-economic situation. Numerous retrospective association studies (24–27) and a few causal inference investigations (28–31) have been reported on SEER's cancer data, making it a unique benchmark resource to assess the actual performance of causal discovery methods on real-life healthcare data.

We present here iMIIC's causal discovery analysis on SEER breast cancer data for the period 2010-2016. There are 407,791 medical records but only 396,179 distinct patients due to multiple breast primary tumors for some patients. Fifty one clinical, socio-economic and outcome variables have been selected for their relevance to breast cancer and for their limited redundancy or missing information, Fig. S1.

The resulting breast cancer network, Fig. 2a, provides an interpretable graphical model including 280 edges, for which most cause-effect relations are either known or can be ruled out based on common or expert knowledge as well as clinical practice. This assessment indicates that about 90% of genuine or putative causal effects inferred by iMIIC are correct, while an additional 8% of cause-effect relations seem plausible, based on clinical and epidemiological knowledge (see Dataset S1). Hence, iMIIC's novel orientation rules lead to only 2% of erroneous causal edges, as compared to about 15% when MIIC's original orientation rules are applied to analyze the same ∼400,000 patient SEER cohort. Besides, none of the predicted genuine causal edges connect pairs of non-cancer-specific variables, such as personal or socio-economic information, that are susceptible to a possible selection bias (13, 14) through breast cancer diagnosis (Box 1). In addition, unmeasured (latent) confounders can be ruled out for genuine causal edges (Box 1) while contributions by measured confounders are estimated as indirect path contributions (see *SI Appendix*, section 6). Yet, other sources of bias in data collection and analysis have been reported on the SEER database (32, 33) (as discussed in the following section). This ∼400,000 patient clinical network is also robust to sub-sampling as it includes 90% of the edges of three networks learned from three random subsets of 100,000 patients, Fig. 2b. In addition, 88% of the edge orientation probabilities are compatible between the three 100,000 patient subset networks and 92% of those are also compatible with the edge orientation probabilities of the full network (see Dataset S1).

Causal interpretation of iMIIC breast cancer network. We now address the clinical and socio-economic interpretation of the SEER breast cancer network inferred by iMIIC, Fig. 2a. We will focus, in particular, on the expected as well as more surprising genuine causal relations uncovered by iMIIC, and will propose interpretations of the counter-intuitive cause-effect predictions in terms of care pathway, therapeutic decisions, patient preferences or socio-economic determinants of healthcare. We present these results from the perspective of different classes of variables and associated subnetworks, starting with the survival subnetwork, then the primary tumor subnetwork, the surgery and subsequent treatment subnetwork, and finally the socio-economic subnetwork.

Survival subnetwork. The full network, Fig. 2a, contains four nodes associated with patient survival status at the end of 2016 and defining a survival subnetwork, that includes all variables directly linked to patient survival status, Fig. 3a. Beyond the vital status of each patient (dead or alive), two additional nodes specify the cause of death, either from breast cancer or from any other cause, and a third continuous variable corresponds to the survival or follow-up delay in months, subjected to the censoring period 2010-2016 of the study. Fig. 3a shows that known factors responsible for the death due to breast cancer are correctly recovered by iMIIC, such as metastatis at diagnosis (overall mortality rate 49.2%), with the worse distant metastases at diagnosis (brain and liver) also retaining direct

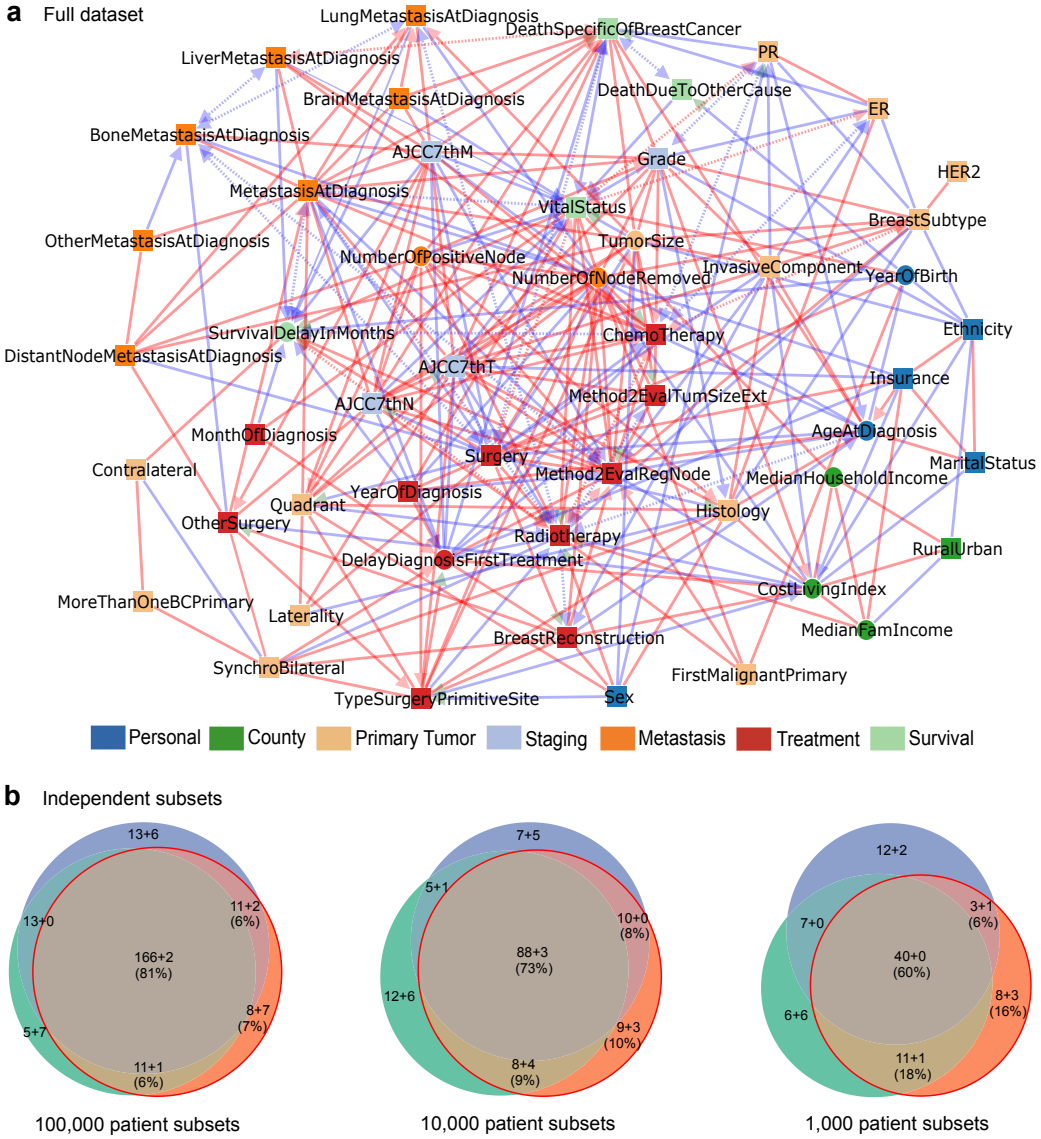


Fig. 2. SEER breast cancer networks inferred by iMIIC. **(a)** The 51 node network inferred by iMIIC from SEER dataset including 396,179 breast cancer patients diagnosed between 2010 and 2016. This skeleton consistent network contains 280 edges and includes 2 contextual variables, Sex and Year of birth. The corresponding orientation consistent network contains 340 edges, Fig. S7. See Dataset S1 for a list and causal nature of each edges predicted by iMIIC. **(b)** Comparisons of networks inferred from three independent sub-samplings of the same size of 100,000, 10,000 or 1,000 patient subsets (from left to right). Number of shared edges (regardless of orientations) in the Euler diagrams are given as a sum $a + b$ where a (resp. b) corresponds to the number of edges included in (resp. absent from) the full dataset network in (a). Percentages refer to the subset network with the median total number of edges (red circle).

links to both Death specific to breast cancer and Vital status, which accounts for their excess mortality rates, *i.e.* brain metastasis (70.5%) and liver metastasis (59.5%). Similarly, the number of metastasis-positive lymph nodes and the staging variables (AJCC7th T, N, and M) are all correctly connected to both death specific to breast cancer and vital status, and not to any other cause of death. By contrast, iMIIC infers causal relations between year of birth and death due to other cause, as well as, year of birth and vital status, as expected. We can also note that the deaths of patients, irrespective of their cause, are rightly predicted to lead to a reduction in their survival delays. Yet, Fig. 3a contains also less intuitive findings. In particular, vital status is robustly inferred to 'cause' radiotherapy, both in the full dataset and in all three 100,000 patient subsets, with 51% of alive patients having undergone radiotherapy against only 27% of dead patients, Fig. 3b. This suggests that early death within the first few months after diagnosis may prevent radiotherapy for some patients who might have otherwise received this treatment, have they lived longer. This short term causal effect between vital status and radiotherapy is consistent with the rapid decline of the survival delay distribution for the first 3-6 months in absence of radiotherapy, Fig. 3c, which corresponds to the typical range of delays for radiotherapy after diagnosis, depending on whether it is performed as second treatment after surgery or as third treatment after both surgery and chemotherapy (34). All in all, this short term causal effect of vital status on radiotherapy outweighs the causally reversed, beneficial effect of radiotherapy on the long term survival of patients. This suggests a strong "immortal time bias" (32)

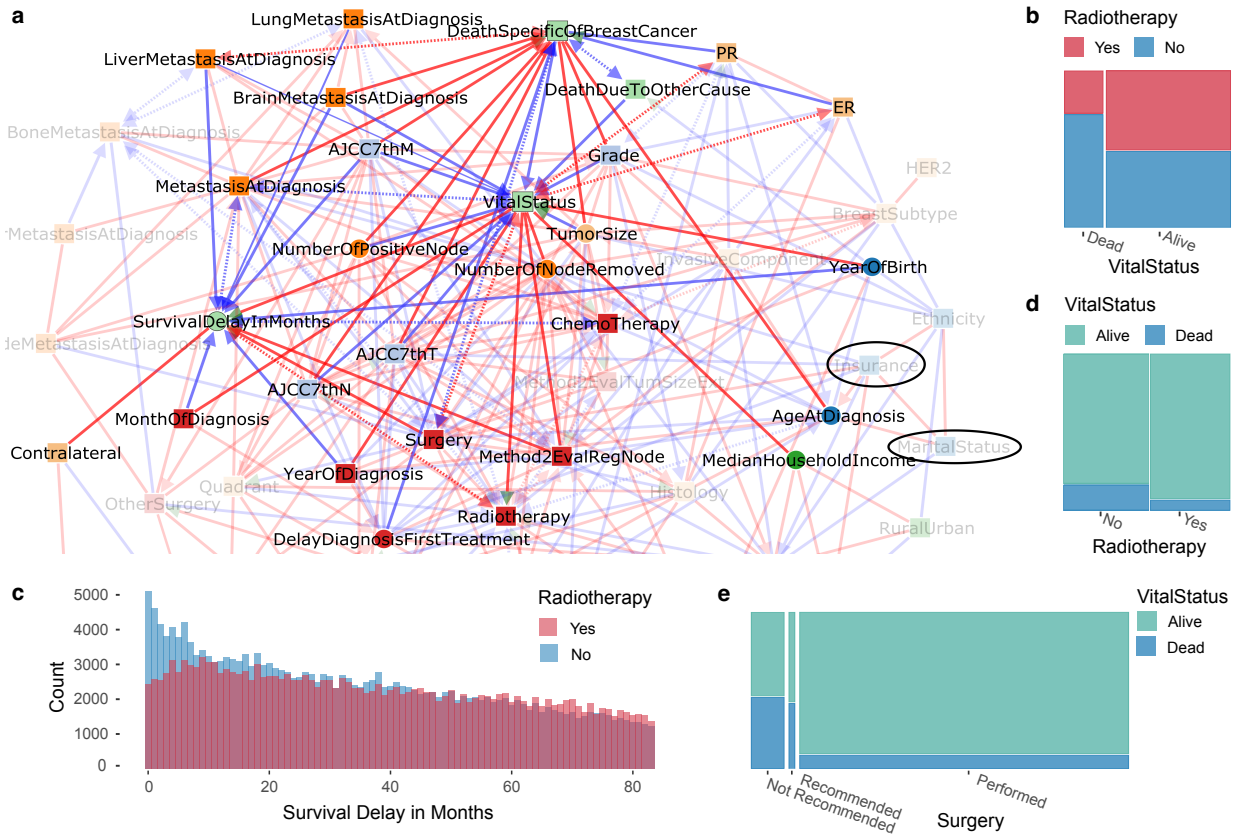


Fig. 3. Survival subnetwork inferred by iMIIC from SEER breast cancer dataset. **(a)** Subnetwork highlighting direct relations with survival variables (VitalStatus, DeathSpecificOfBreastCancer, DeathDueToOtherCause, SurvivalDelayInMonths). The absence of direct links with other variables (such as Insurance and Marital Status highlighted in the network) can be interpreted in terms of indirect path contributions consistent with the network skeleton, see main text and *SI Appendix*, section 6. **(b)** Joint distribution of Radiotherapy and Vital Status highlighting the counter-intuitive causal relation between them, see text. **(c)** Histogram of Survival Delay In Months for patients having received Radiotherapy or not. Each bin represents one month. The early blue peak suggests that a number of patients died within 3 to 6 months after diagnosis, hence, before they could receive Radiotherapy, in agreement with the causal direction predicted in (a). This results in an over-estimated apparent benefit of Radiotherapy in (d), see main text. **(d)** Joint distribution of Vital Status and Radiotherapy. **(e)** Joint distribution of Vital Status and Surgery.

in the apparent benefit of radiotherapy, Fig. 3d, which would need to be corrected with the "landmark method" (32, 35) excluding patients dying within a specified period after surgery, or by emulating a target trial from observational data (36). By contrast, surgery –which is typically performed within 5 to 8 weeks after diagnosis– is found to be the primary cause leading to the prolonged survival delay of patients, as discussed below, Fig. 3e and Fig. 4a.

Finally, we note that a number of variables that have been reported to be associated to survival variables are in fact indirectly rather than directly connected to them. This is, in particular, the case of insurance (37, 38) and marital status (39, 40). The indirect effect of Insurance (with uninsured / medicaid / non-medicaid as categories) on Death due to breast cancer is shown to be indirectly explained through Surgery (50%), ChemoTherapy (14%), MaritalStatus (20%), Radiotherapy (9%), and Breast reconstruction (7%), see Eq. 10 in *SI Appendix*, section 6. Similarly, the indirect effect of marital status (with single / married / separated / divorced / widowed categories) on Death due to breast cancer is shown to be indirectly explained through Surgery (58%), Year of birth (40%), and Ethnicity (2%).

Primary tumor subnetwork. Besides metastasis at diagnosis, the hormone receptor (ER/PR) status and the size of the primary tumor are also found to directly affect the vital prognosis of patients, Fig. S8a. In particular, iMIIC infers that ER status reduces the risk of death due to breast cancer from 17.7% (ER-) to 5.4% (ER+), with a large indirect contribution (82%) from PR status. This is consistent with the ER transcriptional control of PR (41) and a significantly higher mortality rate of ER+/PR- patients (11.8%) than ER+/PR+ patients (4.4%). Likewise, iMIIC infers a number of direct associations between the histology of primary tumors and other variables, Fig. S8a, such as Age at diagnosis (in agreement with early reports (42)) and with synchro bilateral primaries (detected within 6 months of first diagnosis) which are almost twice more likely to occur when lobular carcinoma is present, Fig. S8b. By contrast, no significant association is found with contralateral primary tumors detected more than 6 months after diagnosis, Fig. S8c.

Surgery and subsequent treatment subnetwork. Interestingly, iMIIC also uncovers the central role of Surgery on the precise characterisation of primary tumors, Fig. 4a. For instance, iMIIC uncovers a somewhat unexpected genuine causal link from Surgery to Histology, which reflects that histological types are frequently refined after surgery by the pathologist based on the surgical specimen. This is consistent with a significant increase in histological types including specific tissues after

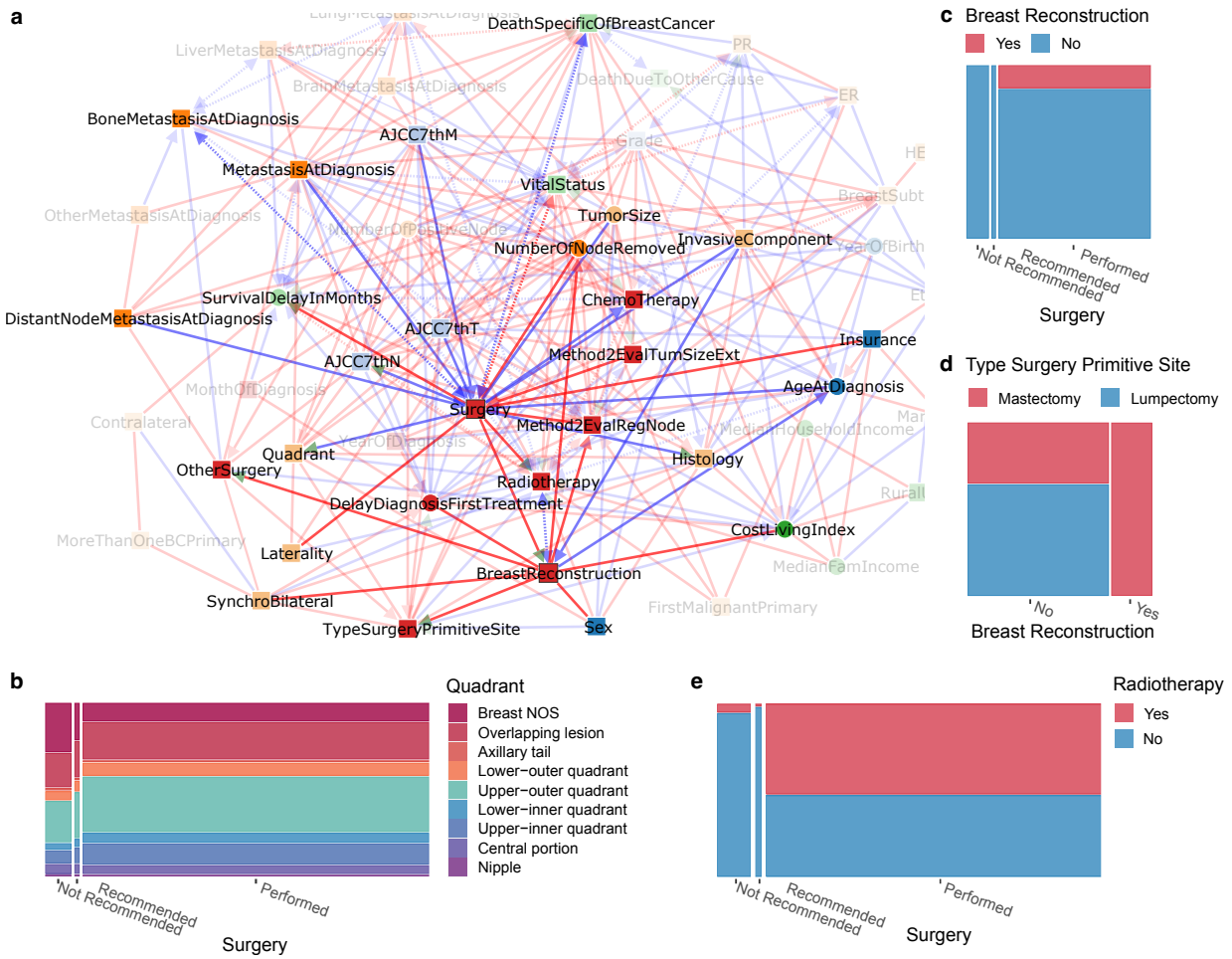


Fig. 4. Surgery and subsequent treatments subnetwork inferred by iMIIC from SEER breast cancer dataset. **(a)** Subnetwork highlighting direct relations with Surgery and Breast Reconstruction. **(b)** Joint distribution of Quadrant and Surgery. **(c)** Joint distribution of Breast Reconstruction and Surgery. **(d)** Joint distribution of Type Surgery Primitive Site and Breast Reconstruction. **(e)** Joint distribution of Radiotherapy and Surgery. See main text for causal interpretation of the role of Surgery on refining primary tumor characterisation and subsequent therapeutic decisions including personal choice of patients.

surgery such as Infiltrating duct mixed with other types of carcinoma (+77% after surgery), Infiltrating duct and lobular carcinoma (+48%), Infiltrating duct carcinoma, NOS (+7.6%), and a corresponding decrease in more generic histological types such as Lobular carcinoma, NOS (-11%), Carcinoma, NOS (-91%), and Adenocarcinoma, NOS (-95%). Similarly, iMIIC rightly infers that the staging variable, AJCC7thN, is usually based on the pathological report following surgery, while not performing surgery (due to the presence of distant metastases at diagnosis or the patient's old age) leads to much more frequent unspecified breast quadrant localisation for primary tumor, Fig. 4a, *i.e.* 30.4% "Breast NOS" when surgery is not recommended *versus* 11.1% when it is performed, Fig. 4b.

Likewise, iMIIC uncovers the central role of Surgery on the therapeutic decisions about subsequent treatments, such as breast reconstruction and radiotherapy, Fig. 4a. While breast reconstruction indeed requires breast surgery, Fig. 4c, iMIIC also correctly infers that the Type of Surgery at the Primary Site (lumpectomy or mastectomy) largely depends on the personal choice of early stage breast cancer patients between breast conservation and reconstruction alternatives, Fig. 4a,d. Similarly, iMIIC rightly infers that radiotherapy is a frequent "consequence" of breast surgery, Fig. 4a, *i.e.* 53% *versus* 4% radiotherapy if surgery is performed or not, Fig. 4e, especially after lumpectomy (75%) to limit the risk of relapse after breast conservation surgery.

Socio-economic subnetwork. The full breast cancer network on Fig. 2a includes four socio-economic variables pertaining to the county of residence of each patient: Median Family Income, Median Household Income, Cost of Living Index and the Rural-Urban population size of each county. These four socio-economic variables actually form a fully connected subgraph (*i.e.* a clique), indicating their strong interdependencies, and are directly connected to a number of other variables, Fig. 5a. Interestingly, Vital Status is only connected to this county variable clique through Median Household Income, which is consistent with earlier reports on the association between life expectancy and incomes (43). By contrast, all other patient specific variables connected to the county clique (such as tumor grade, radiotherapy, breast reconstruction, insurance) have in fact at least one link with Cost of Living Index, highlighting the healthcare system integration into the global economy. In particular, there is a direct association between higher cost of

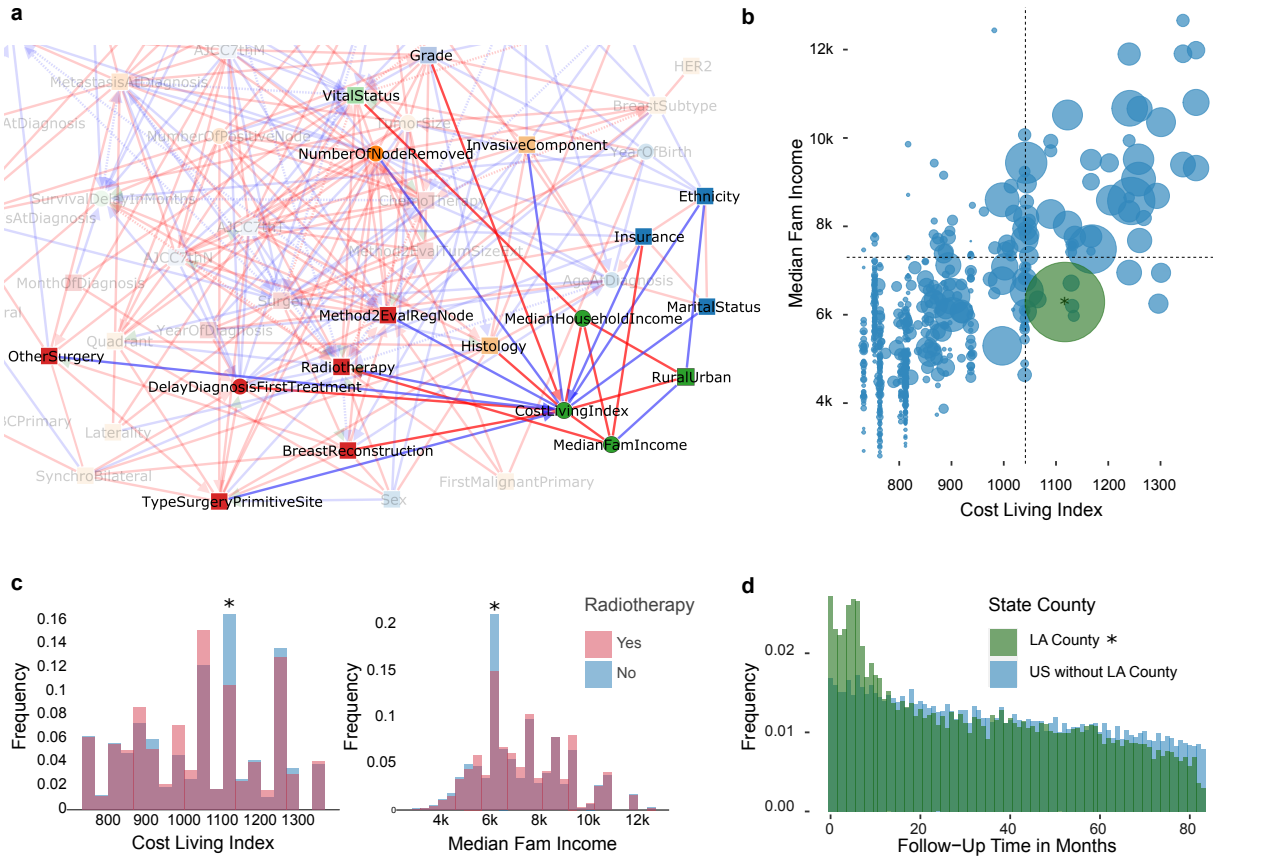


Fig. 5. Socio-economic subnetwork inferred by iMIIC from SEER breast cancer dataset. **(a)** Subnetwork highlighting direct relations with socio-economic county variables (CostLivingIndex, MedianFamIncome, MedianHouseholdIncome, and RuralUrban). **(b)** Bubble plot of the joint distribution of Median Family Income and Cost of Living Index. The bubble area represents the number of patients in that county. Dashed lines correspond to the mean Cost of Living Index and mean Median Family Income. The green bubble with an asterisk corresponds to Los Angeles (L.A.) county which accounts for 10% of the full dataset. **(c)** Histograms of Cost of Living Index and Median Family Income grouped by Radiotherapy. Bins with an asterisk correspond to L.A. county. **(d)** Histograms of Follow-Up Time in Months for L.A. patients and for all other US counties included in SEER.

living and more favorable breast cancer prognosis (*e.g.* fewer invasive components at diagnosis). This is presumably due to better preventive healthcare including easier access to breast cancer screening centers and more comprehensive insurance coverage. Yet, there are also strong disparities between counties, as manifested by the opposite associations of Insurance and Radiotherapy with Median Family Income *versus* Cost of Living Index, Fig. 5a. These intriguing findings can be traced back to Los Angeles (L.A.) county, amounting to about 10% of the whole dataset, which presents a lower than average median family income (29-38% percentile range) despite a higher than average cost of living index (58-67% percentile range), Fig. 5b. This must have led to an exacerbated financial burden for many of the 39,089 breast cancer patients diagnosed in L.A. county between 2010 and 2016. Although 18% of these patients benefited from medicaid insurance (as compared to 10% in the whole dataset), many had to opt for affordable but limited private insurance including significant co-payment policies or even to become uninsured especially before the application of the Affordable Care Act in January 2014 (3.4% uninsured in 2013 against 1.5% in 2014). As a result, many L.A. patients appear to have renounced to undergo expensive treatments. In particular, only 32.6% of patients underwent radiotherapy in L.A. as compared to 50% of patients nationwide excluding L.A. county, Fig. 5c, which can only be partly accounted for by county differences in under-reported radiotherapy of outpatients (32, 33). Moreover, an estimated 7% of L.A. patients even appear to have dropped out of therapy or moved to a different county not included in SEER database (against 1.5% nationwide, excluding L.A. county), based on the rapidly decreasing follow-up time distribution in L.A. as compared to the rest of the dataset, Fig. 5d. This corresponds to the fraction of patients having had their last medical contact less than a year after diagnosis and more than a year before the end of this study in December 2016.

Discussion

Nationwide healthcare data, such as the SEER breast cancer dataset analyzed here, are especially interesting from a methodological point of view; they provide real-life benchmark datasets, which can help assess the reliability of causal discovery methods on real-life data, as most cause-effect predictions can be validated or dismissed, based on expert knowledge, clinical practice or possible data collection and selection biases. Besides, the interpretability of Machine Learning methods is particularly relevant for applications on clinical data, for which Artificial Intelligence assisted recommendations can hardly rely on black box classifiers only and need to be explainable in terms of intelligible rationales to medical practitioners. Yet,

beyond clinical data, causal discovery methods have the potential to become essential Machine learning approaches to interpret diverse observational data in a wide range of domains, for which systematic perturbation experiments are not available due to practical, cost or ethical reasons. In particular, causal discovery can guide biological research by predicting the causal effects of specific interventions (44), such as gene expression or gene silencing, which can then be probed by targeted siRNA, gene knock-out or CRISPR-based editing experiments.

In the context of SEER's breast cancer dataset, iMIIC uncovers many expected causal relations, such as the adverse consequence of metastasis and the protecting effect of ER+ status on death due to breast cancer, or the fact that year of birth is the primary reason for death due to other causes by the end of the study. On the other hand, the effects of insurance coverage or marital status, which have been reported to reduce the risk of death due to breast cancer, are found to be entirely indirect and mainly mediated by treatments (60-80%), notably, surgery (>50%). In fact, surgery appears as the cornerstone of breast cancer therapy by first helping refine histological types, then guide therapeutic decisions on radiotherapy and breast reconstruction and ultimately prolong the survival delays of patients. Yet, iMIIC also correctly infers that the type of surgery (lumpectomy or mastectomy) at the primary site largely depends on the personal choice of early stage breast cancer patients between breast conservation or reconstruction alternatives. By contrast, other treatments, such as radiotherapy and chemotherapy, seem to have less decisive impacts on breast cancer outcome, which might be due in part to some under-reported treatment information in the SEER database (32, 33). Radiotherapy even appears to be a consequence, not a cause, of vital status, suggesting that early death within the first few months after diagnosis may prevent radiotherapy for some patients who might have otherwise received this treatment, have they lived longer. Finally, iMIIC recovers direct associations between socio-economic county variables (such as median family income and cost of living index) and patient specific variables (such as tumor grade, radiotherapy, breast reconstruction, insurance), highlighting the healthcare system integration into the global economy. While higher costs of living are on average associated to more favorable cancer prognosis, presumably due to better preventive healthcare and more comprehensive insurance coverage, iMIIC also uncovers large disparities between family income and cost of living indices across counties (*e.g.* for L.A. county), leading to exacerbated financial burden with patients giving up expensive treatments or even dropping out of treatment.

In summary, iMIIC is a general causal discovery method, which uncovers direct and possibly causal relations as well as network consistent indirect effects for a broad range of biological and clinical data. iMIIC can handle heterogeneous data (*i.e.* combining categorical and continuous variables), data with unobserved latent variables (*i.e.* with unmeasured confounders), as well as, variables with missing data, that are ubiquitous in many real-life applications. Importantly, iMIIC is fully unsupervised and does not need preconceived hypothesis nor expert knowledge. In particular, iMIIC automatically adjusts for measured confounders (in the form of indirect contributions) and distinguishes genuine causes from putative and latent causal effects by either ruling out or highlighting the effect of unmeasured confounders for each causal edge (Box 1). While iMIIC is not immune to possible data collection and selection biases, which can affect observational data, it is based on a robust information theoretic framework, making it particularly reliable to interpret challenging types of data, such as multimodal data integrated from different sources (*e.g.* clinical, personal, socio-economic data, as demonstrated here and in (10, 45)) or different experimental techniques (*e.g.* single cell transcriptomics (8, 44, 46) and imaging data (10)). In principle, iMIIC could be applied to a wide range of other domains to uncover causal relations and quantify indirect contributions when only observational data is available. With the advent of virtually unlimited datasets in many data science domains, scalable causal discovery methods are much needed and we believe that iMIIC can bring unique insights based on causal interpretation across a range of research fields.

Materials and Methods

SEER-like dataset generation. SEER-like synthetic datasets were generated using network structures inferred from 10,000 patient subsets of the full SEER dataset of breast cancer patients, to allow for comparison with other causal discovery methods, as detailed below. Random network skeletons of similar SEER-like degree distributions with additional ± 2 connection variability at each node were first obtained using a Monte Carlo graph generation algorithm (47). These skeletons were subsequently oriented to obtain Directed Acyclic Graphs using a random ordering of their nodes and assigning various proportions of discrete *versus* continuous variables. The marginal distributions of variables without parents were chosen to resemble typical SEER-like marginal distributions, Fig. S1, and the other variables were simulated using mixed-type structural equation models (SEMs) (10), see *e.g.* Fig. S2. For each discrete node proportion (decile steps), 25 benchmark networks were obtained and used to generate 100,000 samples each.

Causal discovery scores. For evaluation purposes, network reconstruction was treated as a binary classification task and classical performance measures, Precision, Recall and F-score, were computed to evaluate (*i*) skeleton, (*ii*) completed partially directed acyclic graph (CPDAG) and (*iii*) oriented-edge subgraph reconstructions. CPDAG scores use the same metrics as skeleton scores but rating as "false positive" the erroneous orientation of non-oriented edges in the CPDAG and the non-orientation or opposite orientation of oriented edges in the CPDAG. However, these errors are not equivalent from a causal discovery perspective. Hence, we introduced oriented-edge subgraph scores, that are restricted to the subgraphs containing only oriented edges in the theoretical CPDAG *versus* the inferred graph. These oriented-edge scores, highlighted in the benchmark comparisons, are designed to specifically assess the method performance on causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.

Benchmarked causal discovery methods. Five causal discovery methods able to analyze mixed-type datasets have been compared over SEER-like generated datasets:

- *Interpretable MIIC (iMIIC)* was run with default parameters for all settings.
- *Original MIIC* (9, 10) was run with default parameters for all settings (Fig. 1c and Fig. S3).
- *PC*(18) from the *pcalg* package (19) was run with the stable option (48) and either majority rule (48) (Fig. S3) or conservative rule (20) (Fig. 1d and Fig. S4) for orientations. The "ci.test" function from the *bnlearn* package (49) was used as independence test for mixed-type data (with either "mi-cg" option for discrete against continuous variables, "mi" for discrete against discrete variables or "mi-g" for continuous against continuous variables) and the threshold for significance testing was set to the default $\alpha = 0.01$.

- *causalMGM* (21) was run with the *rCausalMGM* R package. The initial graph was computed using the mgm() function with each of the 3 lambda parameters equal to 0.05 and the orientations were then obtained with the pcMax() function with default $\alpha = 0.01$ parameter.
- *MXM* (22), a mixed-PC constraint-based method, was run using the *MXM* R package. The graph was obtained using the pc.skel() function for skeleton with the "comb.mm" independence test and the default $\alpha = 0.01$ threshold for significance testing and with the pc.or() function for orientations.

Computation time. Benchmarks were stopped when the average computation time of a method reached 1 hour per network with high proportion of continuous variables (resp. about 10 minutes per network with low proportion of continuous variables), corresponding to a maximum running time of about 115h for the 250 generated networks at each sample size.

Benchmark results. The performance of iMIIC has been benchmarked against state-of-the-art constraint-based methods: PC, causalMGM and MXM, on SEER-like benchmark datasets with different proportions of discrete variables, Fig. 1d and Figs. S4-S6. Results for datasets with 80% discrete variables, corresponding to the actual proportion in the real-life SEER breast cancer dataset, are discussed in the main text. Similarly, for larger proportions of continuous variables, Fig. 1d and Figs. S4-S6 demonstrate that iMIIC greatly outperforms the reliability and sensitivity of predicted orientations against state-of-the-art constraint-based methods. For instance, for SEER-like benchmark datasets with only 20% of discrete variables, iMIIC already reaches 81% (resp. 64%) in precision (resp. F-score), for $N = 10^3$, against 53% (29%) for conservative PC, 50% (40%) for causalMGM and 29% (25%) for MXM. For $N = 10^4$, iMIIC reaches 88% (78%) in precision (F-score), against about 60% (45%) for conservative PC, 52% (50%) for causalMGM and 22% (28%) for MXM. Finally, iMIIC reaches 86% (81%) for $N = 10^5$, which is beyond the sample size attainable by other methods.

Data, materials and software availability. The dataset of breast cancer patients was obtained from the Surveillance, Epidemiology and End Results program, which can be accessed at https://seer.cancer.gov/seertrack/data/request/. Causal discovery using iMIIC was performed on the open access server https://miic.curie.fr or running the R package available at https://github.com/miicTeam/miic_R_package. Other R packages used for benchmark comparisons are available at https://r-forge.r-project.org/projects/pcalg, https://cran.r-project.org/web/packages/bnlearn, https://github.com/tyler-lovelace1/rCausalMGM and https://cran.r-project.org/web/packages/MXM.

ACKNOWLEDGMENTS. We would like to thank Irène Buvat, Laura Cantini, Michèle Sebag, Jean-Christophe Thalabard and Nathalie Vialaneix for helpful discussions and comments on a first version of the manuscript.

1. P Spirtes, CN Glymour, R Scheines, D Heckerman, *Causation, prediction, and search*. (MIT press), (2000).
2. J Pearl, *Causality*. (Cambridge university press), (2009).
3. J Runge, P Nowack, M Kretschmer, S Flaxman, D Sejdinovic, Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.* **5** (2019).
4. J Runge, et al., Inferring causation from time series in earth system sciences. *Nat. Commun.* **10**, 2553 (2019).
5. D Heckerman, D Geiger, DM Chickering, Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Mach. Learn.* **20**, 197–243 (1995).
6. J Friedman, T Hastie, R Tibshirani, Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).
7. J Zhang, On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* **172**, 1873–1896 (2008).
8. L Verny, N Sella, S Affeldt, PP Singh, H Isambert, Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.* **13**, e1005662 (2017).
9. N Sella, L Verny, G Uguzzoni, S Affeldt, H Isambert, Miic online: a web server to reconstruct causal or non-causal networks from non-perturbative data. *Bioinformatics* **34**, 2311–2313 (2018).
10. V Cabeli, et al., Learning clinical networks from medical records based on information estimates in mixed-type data. *PLOS Comput. Biol.* **16**, e1007866 (2020).
11. N Howlader, et al., Seer cancer statistics review in *SEER Cancer Statistics Review*. (National Cancer Institute), pp. 1975–2016 (2018).
12. J Peters, JM Mooij, D Janzing, B Schölkopf, Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.* **15**, 2009–2053 (2014).
13. DL Sackett, Bias in analytic research. *J. Chronic Dis.* **32**, 51–63 (1979).
14. MA Hernán, S Hernández-Díaz, JM Robins, A structural approach to selection bias. *Epidemiology* **15**, 615–625 (2004).
15. TM Cover, JA Thomas, *Elements of Information Theory*. (Wiley), 2nd edition, (2006).
16. V Cabeli, H Li, M da Câmara Ribeiro-Dantas, F Simon, H Isambert, Reliable causal discovery based on mutual information supremum principle for finite datasets in *Paper presented at WHY21 workshop, 35rd Conference on Neural Information Processing Systems*. (NeurIPS), (2021).
17. H Li, V Cabeli, N Sella, H Isambert, Constraint-based causal structure learning with consistent separating sets. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **32** (2019).
18. P Spirtes, C Glymour, An algorithm for fast recovery of sparse causal graphs. *Soc. Sci. Comput. Rev.* **9**, 62–72 (1991).
19. M Kalisch, M Mächler, D Colombo, MH Maathuis, P Bühlmann, Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* **47**, 1–26 (2012).
20. J Ramsey, P Spirtes, J Zhang, Adjacency-faithfulness and conservative causal inference in *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, UAI. (AUAI Press, Oregon, USA), pp. 401–408 (2006).
21. AJ Sedgewick, et al., Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. *Bioinformatics* **35**, 1204–1212 (2018).
22. M Tsagris, G Borboudakis, V Lagani, I Tsamardinos, Constraint-based causal discovery with mixed data. *Int. J. Data Sci. Anal.* **6**, 19–30 (2018).
23. N Harbeck, et al., Breast cancer. *Nat. Rev. Dis. Primers* **5**, 6 (2019).
24. AM Alaa, D Gurdasani, AL Harris, J Rashbass, M van der Schaar, Machine learning to guide the use of adjuvant therapies for breast cancer. *Nat. Mach. Intell.* **3**, 716–726 (2021).
25. C Lee, et al., Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the surveillance, epidemiology, and end results (SEER) database. *The Lancet Digit. Heal.* **3**, e158–e165 (2021).
26. G Mendiratta, et al., Cancer gene mutation frequencies for the U.S. population. *Nat. Commun.* **12**, 5961 (2021).
27. HG Welch, PC Prorok, AJ O'Malley, BS Kramer, Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. *New Engl. J. Medicine* **375**, 1438–1447 (2016).
28. MS Leapman, et al., Mediators of Racial Disparity in the Use of Prostate Magnetic Resonance Imaging Among Patients With Prostate Cancer. *JAMA Oncol. Publ. online March 03* (2022).
29. LC Petito, et al., Estimates of overall survival in patients with cancer receiving different treatment regimens. *JAMA Netw. Open* **3**, e200452 (2020).
30. RC Nethery, Y Yang, AJ Brown, F Dominici, A causal inference framework for cancer cluster investigations using publicly available data. *J. Royal Stat. Soc. Ser. A (Statistics Soc.* **183**, 1253–1272 (2020).
31. L Wang, Mining causal relationships among clinical variables for cancer diagnosis based on bayesian analysis. *BioData Min.* **8**, 13 (2015).
32. HS Park, S Lloyd, RH Decker, LD Wilson, JB Yu, Limitations and biases of the surveillance, epidemiology, and end results database. *Curr. Probl. Cancer* **36**, 216–224 (2012).
33. R Jagsi, et al., Underascertainment of radiotherapy receipt in surveillance, epidemiology, and end results registry data. *Cancer* **118**, 333–341 (2011).
34. SY Chen, et al., Timing of chemotherapy and radiotherapy following breast-conserving surgery for early-stage breast cancer: A retrospective analysis. *Front. Oncol.* **10**, 571390 (2020).
35. JR Anderson, KC Cain, RD Gelber, Analysis of survival by tumor response. *J. Clin. Oncol.* **1**, 710–719 (1983).
36. MA Hernán, JM Robins, Using big data to emulate a target trial when a randomized trial is not available: Table 1. *Am. J. Epidemiol.* **183**, 758–764 (2016).
37. X Han, KR Yabroff, E Ward, OW Brawley, A Jemal, Comparison of insurance status and diagnosis stage among patients with newly diagnosed cancer before vs after implementation of the patient protection and affordable care act. *JAMA Oncol.* **4**, 1713 (2018).
38. T Ermer, et al., Understanding the implications of medicaid expansion for cancer care in the US. *JAMA Oncol.* **8**, 139 (2022).
39. L Hinyard, LS Wirth, JM Clancy, T Schwartz, The effect of marital status on breast cancer-related outcomes in women under 65: A seer database analysis. *The Breast* **32**, 13–17 (2017).
40. Z Zhai, et al., Effects of marital status on breast cancer survival by age, race, and hormone receptor status: A population-based study. *Cancer medicine* **8**, 4906–4917 (2019).
41. J Bonéy-Montoya, YS Ziegler, CD Curtis, JA Montoya, AM Nardulli, Long-range transcriptional control of progesterone receptor gene expression. *Mol Endocrinol* **24**, 346–358 (2010).
42. C Fisher, et al., Histopathology of breast cancer in relation to age. *Br. journal cancer* **75**, 593–596 (1997).
43. R Chetty, et al., The association between income and life expectancy in the united states, 2001-2014. *JAMA* **315**, 1750 (2016).
44. C Desterke, et al., Inferring Gene Networks in Bone Marrow Hematopoietic Stem Cell-Supporting Stromal Niche Populations. *iScience* **23**, 101222 (2020).
45. N Sella, et al., Interactive exploration of a global clinical network from a large breast cancer cohort. *NPJ Digit. Med.* **5**, 113 (2022).
46. S Affeldt, L Verny, H Isambert, 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinforma.* **17**, 12 (2016).
47. F Viger, M Latapy, Efficient and simple generation of random simple connected graphs with prescribed degree sequence in *Lecture Notes in Computer Science*. (Springer Berlin Heidelberg), pp. 440–449 (2005).
48. D Colombo, MH Maathuis, Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15**, 3741–3782 (2014).
49. M Scutari, Learning Bayesian Networks with the bnlearn R Package. *J. Stat. Softw.* **35**, 1–22 (2010).

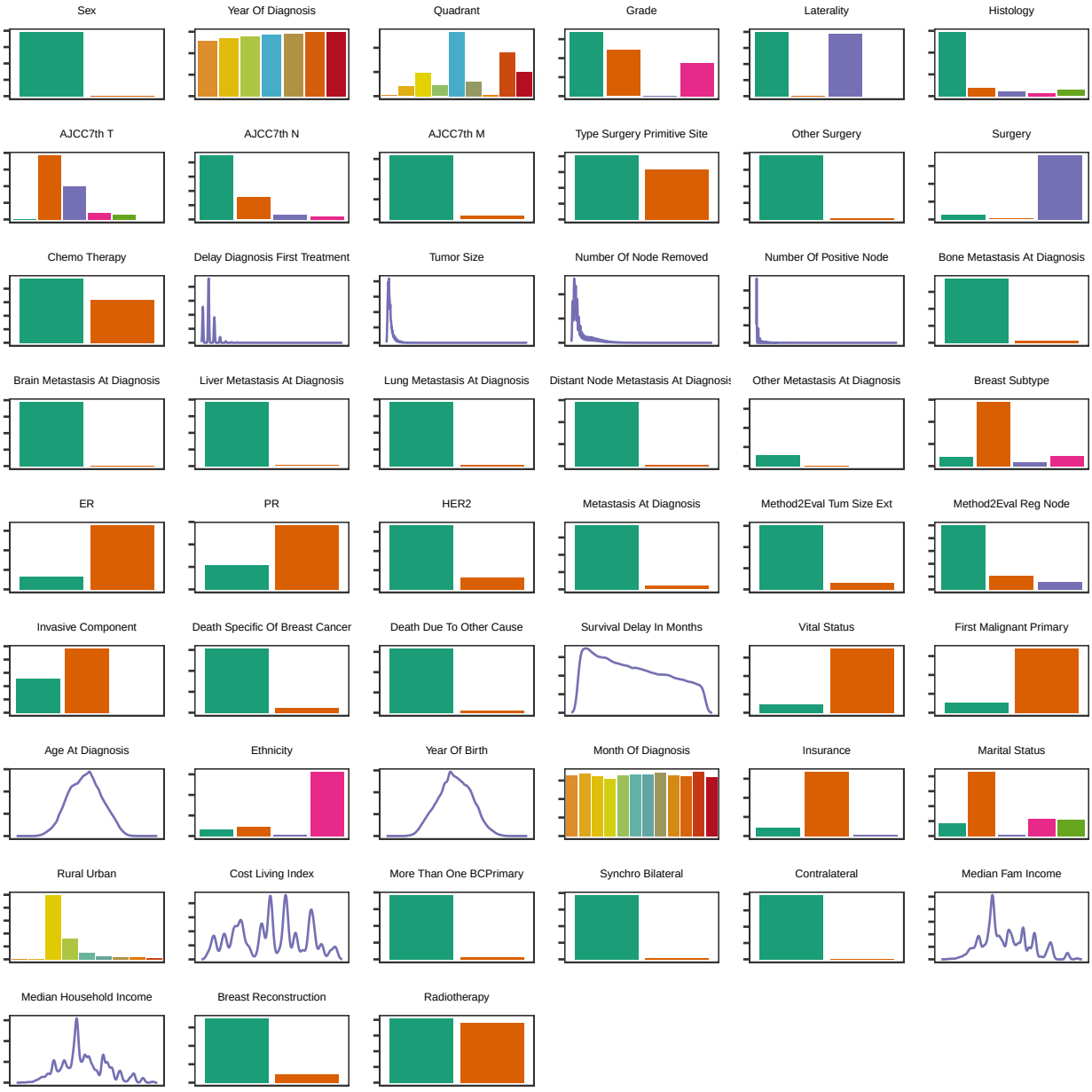


Fig. S1. Distributions of the 51 SEER variables selected for breast cancer. There are 407,791 breast cancer records in SEER for the period 2010-2016, but only 396,179 distinct patients due to multiple breast primary tumors for some patients. For each patient, we selected the first breast primary tumor recorded in SEER and indicated the total number of breast cancer primaries during the 2010-2016 period in the variable *MoreThanOneBCPrimary*. *SynchroBilateral* was also engineered to identify patients who had tumors in both breasts diagnosed within less than 180 days of each other, while *Contralateral* identifies patients who had a subsequent tumor in the other breast diagnosed more than 180 days after the first breast tumor primary. Some categorical variables had some of their categories merged, either because these categories had the same general meaning or because they were too rare amongst patients (*i.e.* $<$0.1% of patients excluding those with missing data for the considered variable). These variables include *Ethnicity, TypeSurgeryPrimitiveSite, Surgery, OtherSurgery, OtherMetastasisAtDiagnosis, Insurance and Histology*. Hence, categories recorded in less than 0.1% of patients were merged and renamed to 'Other'. *BreastReconstruction* was engineered based on *TypeSurgeryPrimitiveSite* (*i.e.* SEER surgery code ranges 43-49, 53-59, 63-69, and 73-75 were assigned 'Yes', while other surgery codes were assigned 'No'). *Radiotherapy* was created from *Radiation sequence with surgery*, that has much fewer missing data (0.05%) than the original *Radiation* variable (49%). *TumorSize* merges two distinct variables that contained tumor sizes for years 2004-2015 and 2016+, respectively. Likewise, the largely missing 2016 information for the *MetastasisAtDiagnosis* variable was recovered based on information contained in specific metastasis variables (*i.e. BoneMetastasisAtDiagnosis, LungMetastasisAtDiagnosis, LiverMetastasisAtDiagnosis, BrainMetastasisAtDiagnosis, OtherMetastasisAtDiagnosis*). Finally, *MedianFamIncome* and *MedianHouseHoldIncome* are the average of these continuous variables over the periods 2007-2011, 2008-2012, 2009-2013, 2010-2014, 2011-2015, and 2012-2016.

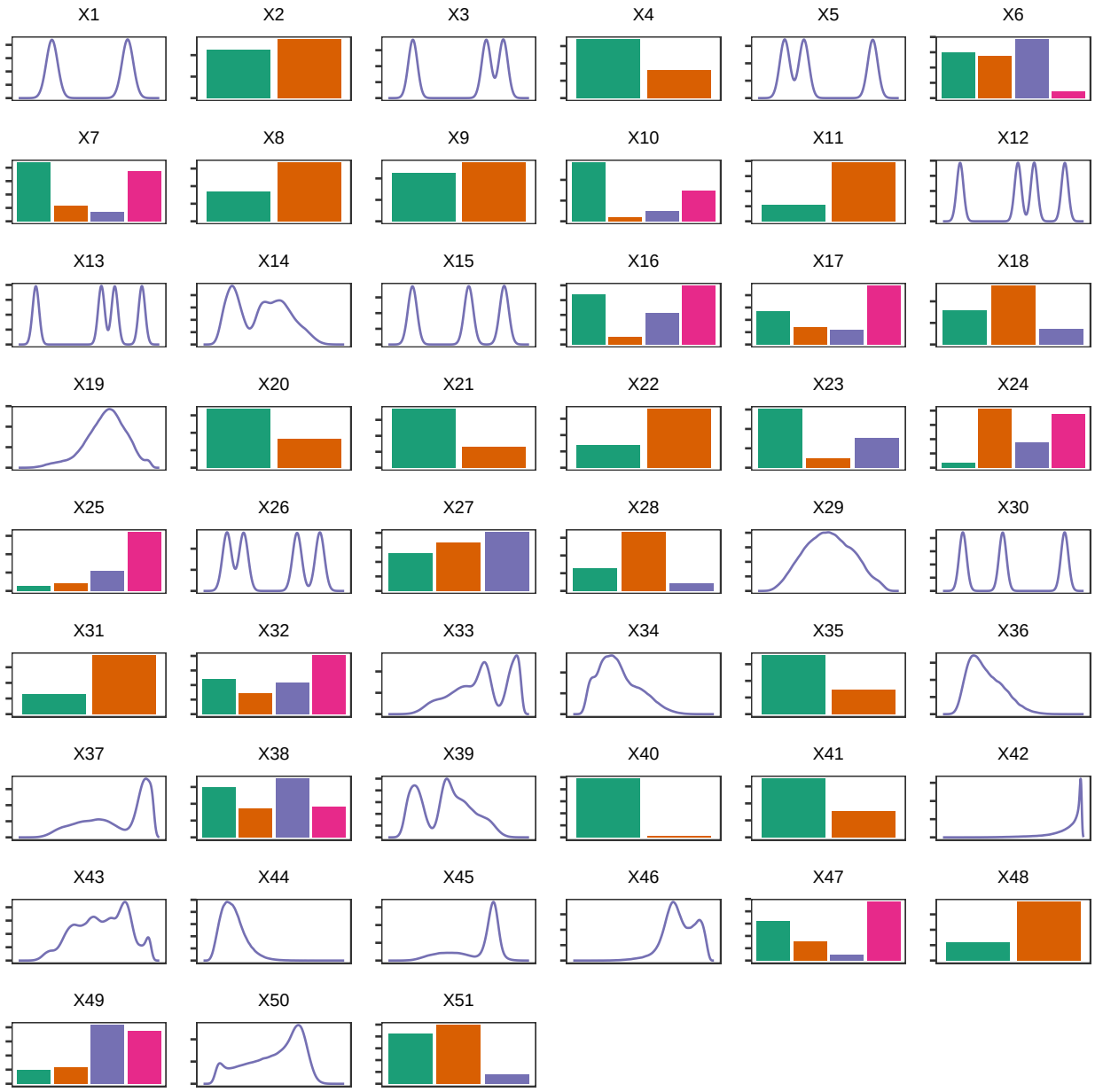


Fig. S2. Example of simulated SEER-like dataset. Example of marginal distributions of simulated SEER-like datasets (including about 60% of discrete variables here) obtained using mixed-type structural equation models (SEMs) (10), see Data generation and benchmarks in Materials and Methods.

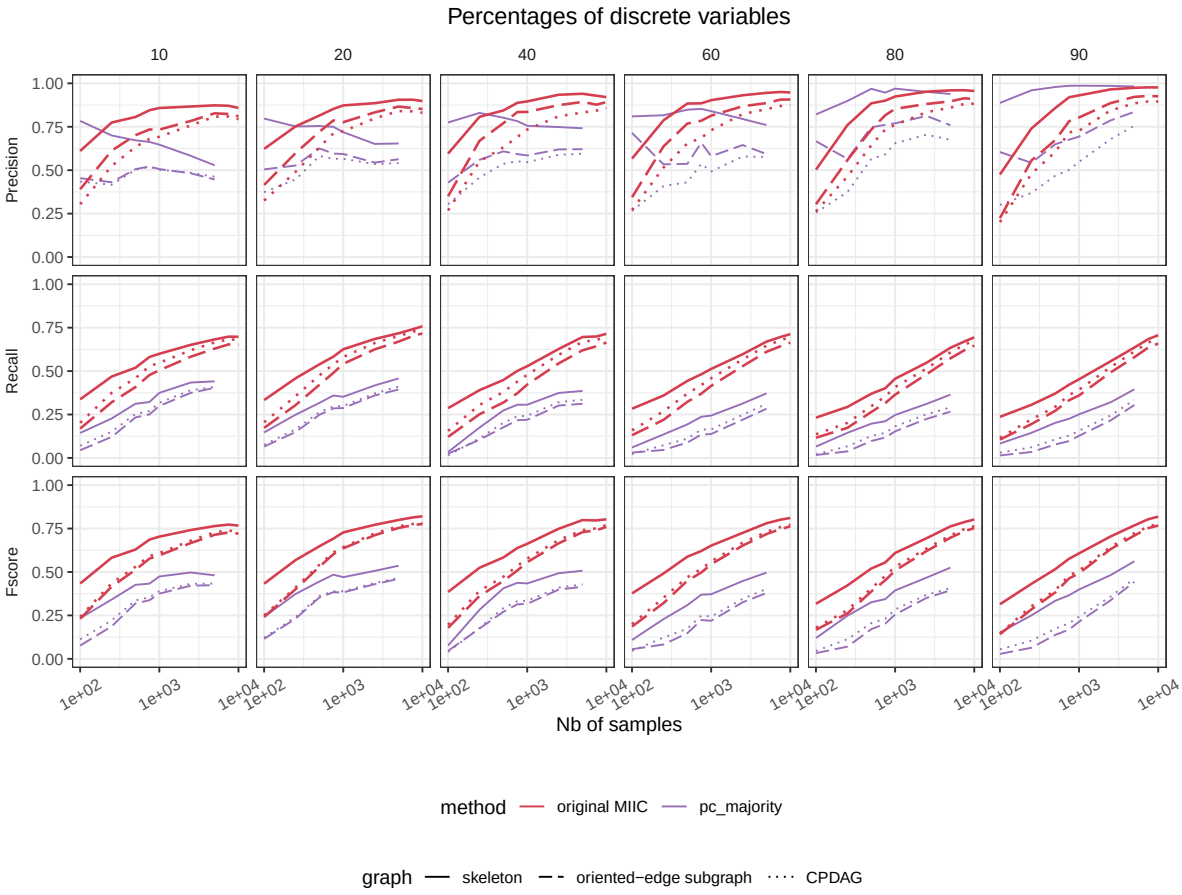


Fig. S3. Original MIIC *versus* PC on SEER-like benchmarks. See parameter settings in Data generation and benchmarks in Materials and Methods. Oriented-edge subgraph scores (dashed lines) are restricted to the subgraphs containing only oriented edges in the theoretical CPDAG *versus* the inferred graph. These oriented-edge scores are designed to specifically assess the method performance on causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.

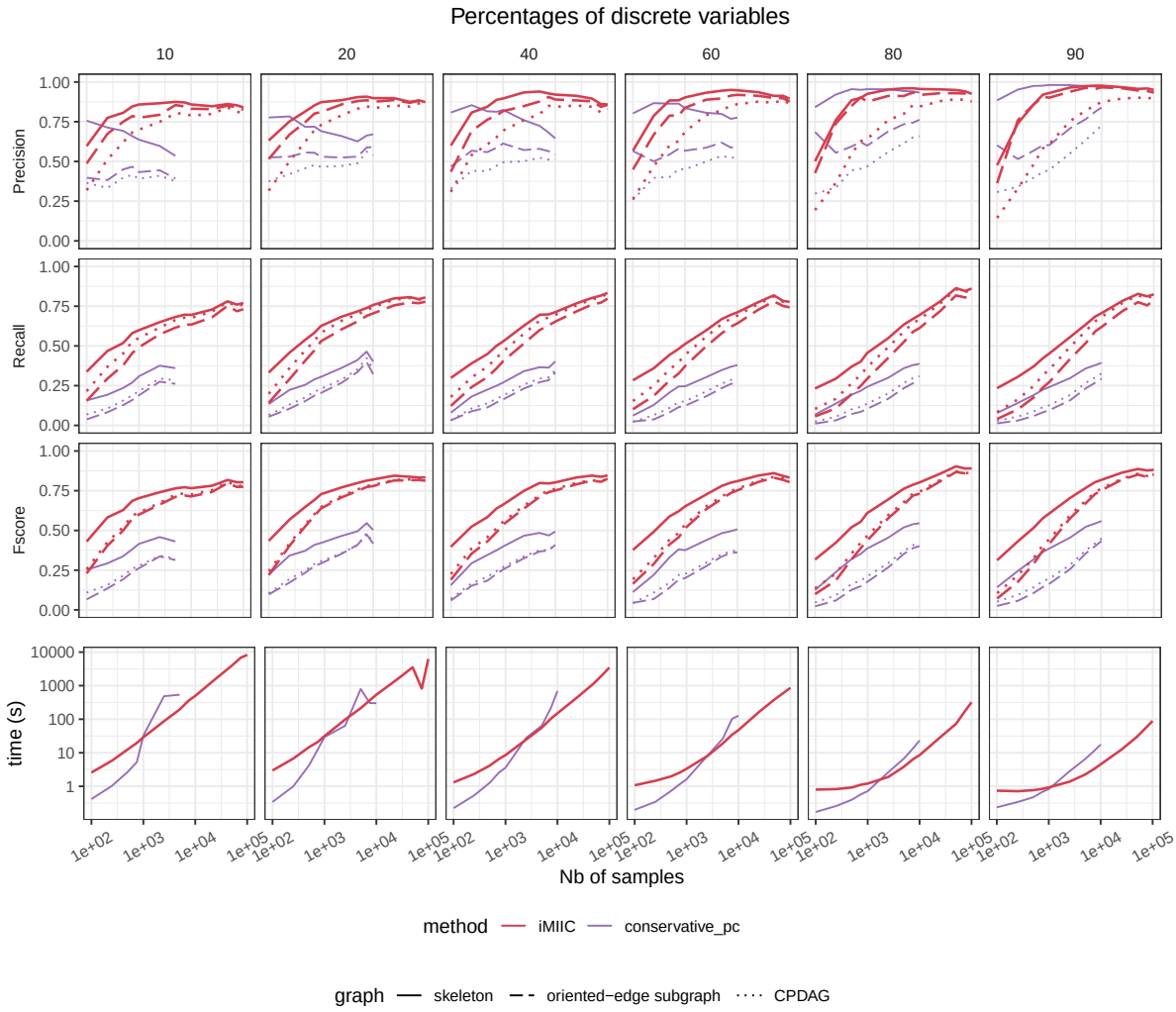


Fig. S4. iMIIC *versus* PC on SEER-like benchmarks. See parameter settings in Data generation and benchmarks in Materials and Methods. Oriented-edge subgraph scores (dashed lines) are restricted to the subgraphs containing only oriented edges in the theoretical CPDAG *versus* the inferred graph. These oriented-edge scores are designed to specifically assess the method performance on causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.

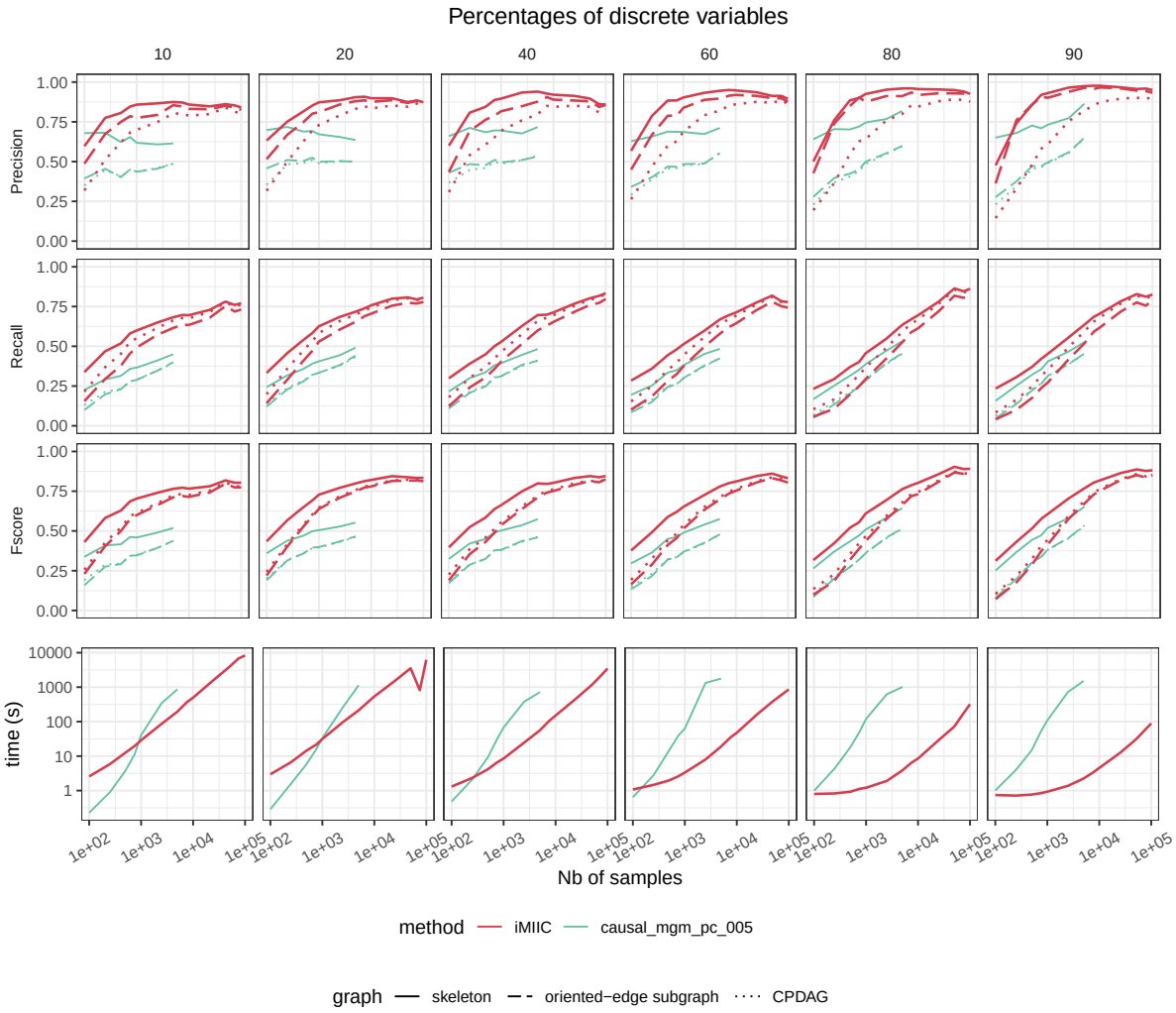


Fig. S5. iMIIC *versus* causalMGM on SEER-like benchmarks. See parameter settings in Data generation and benchmarks in Materials and Methods. Oriented-edge subgraph scores (dashed lines) are restricted to the subgraphs containing only oriented edges in the theoretical CPDAG *versus* the inferred graph. These oriented-edge scores are designed to specifically assess the method performance on causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.

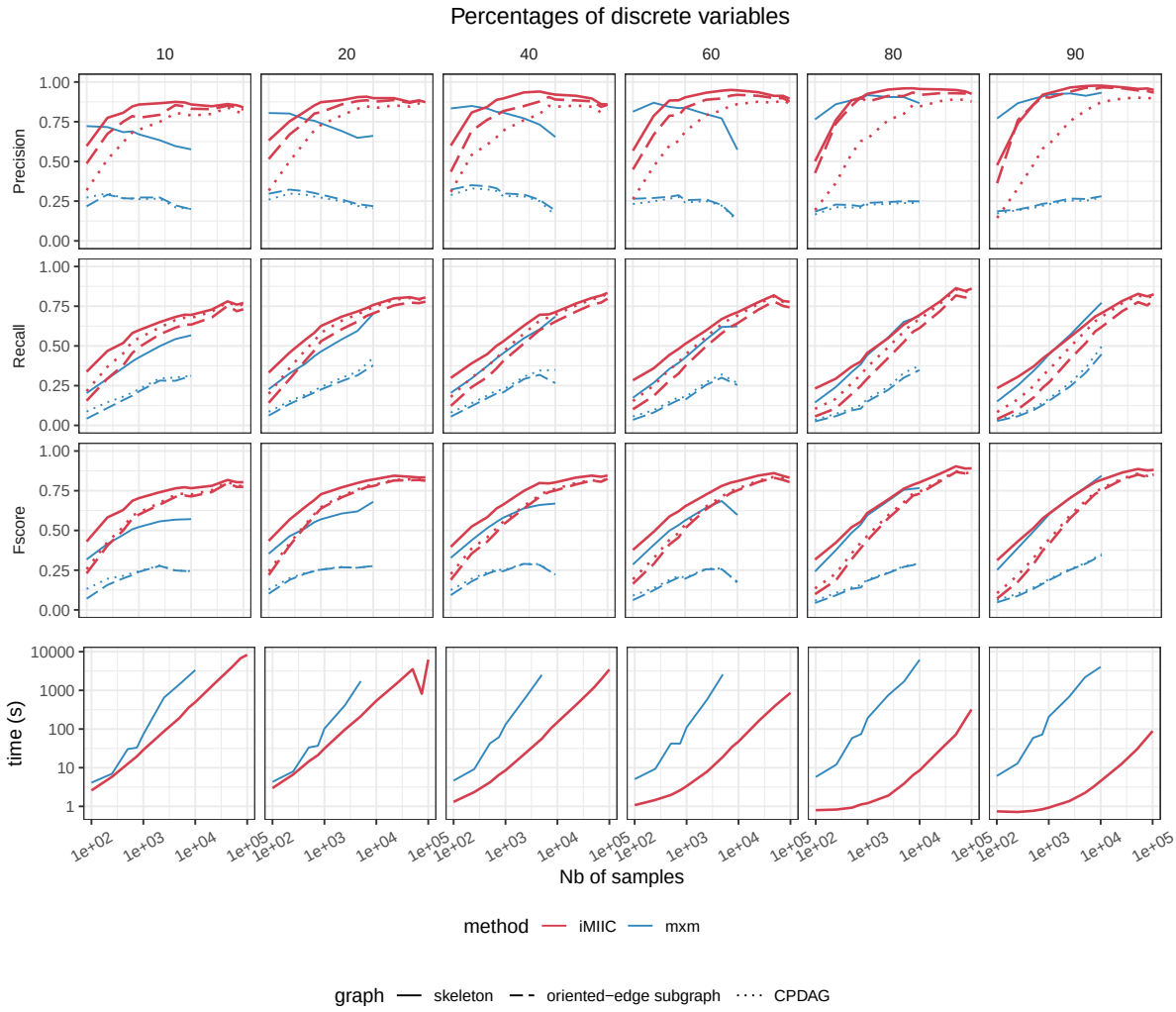


Fig. S6. iMIIC *versus* MXM on SEER-like benchmarks. See parameter settings in Data generation and benchmarks in Materials and Methods. Oriented-edge subgraph scores (dashed lines) are restricted to the subgraphs containing only oriented edges in the theoretical CPDAG *versus* the inferred graph. These oriented-edge scores are designed to specifically assess the method performance on causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.

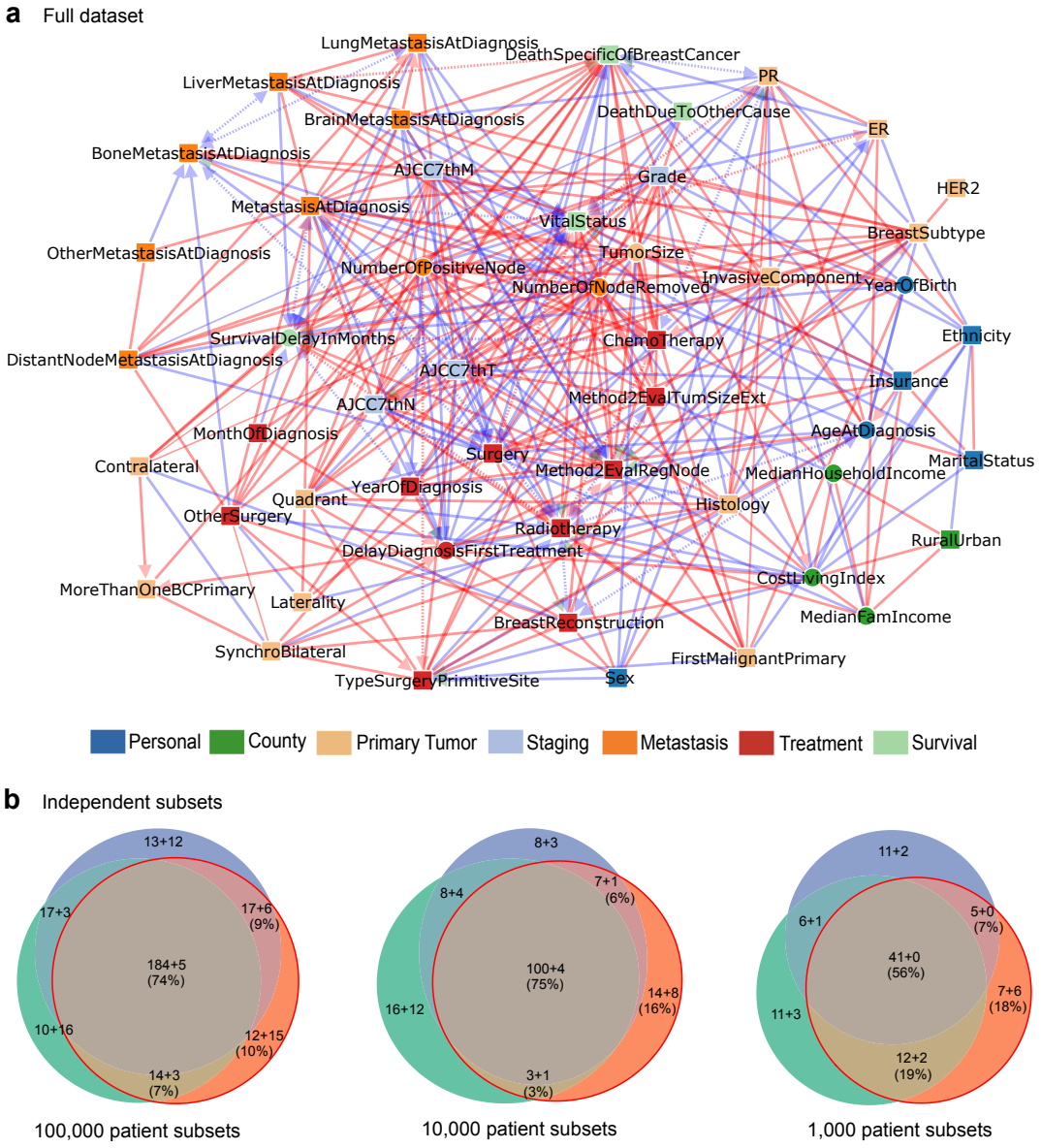


Fig. S7. SEER breast cancer orientation consistent networks inferred by iMIIC. **(a)** The 51 node network inferred by iMIIC from SEER dataset containing 396,179 breast cancer patients diagnosed between 2010 and 2016. This orientation consistent network contains 340 edges and includes 2 contextual variables, Sex and Year of birth. See Dataset S1 for a list and causal nature of each edges predicted by iMIIC. **(b)** Comparisons of networks inferred from three independent sub-samplings of the same size of 100,000, 10,000 or 1,000 patient subsets (from left to right). Number of shared edges (regardless of orientations) in the Euler diagrams are given as a sum $a + b$ where a (resp. b) corresponds to the number of edges included in (resp. absent from) the full dataset network in (a). Percentages refer to the subset network with the median total number of edges (red circle).

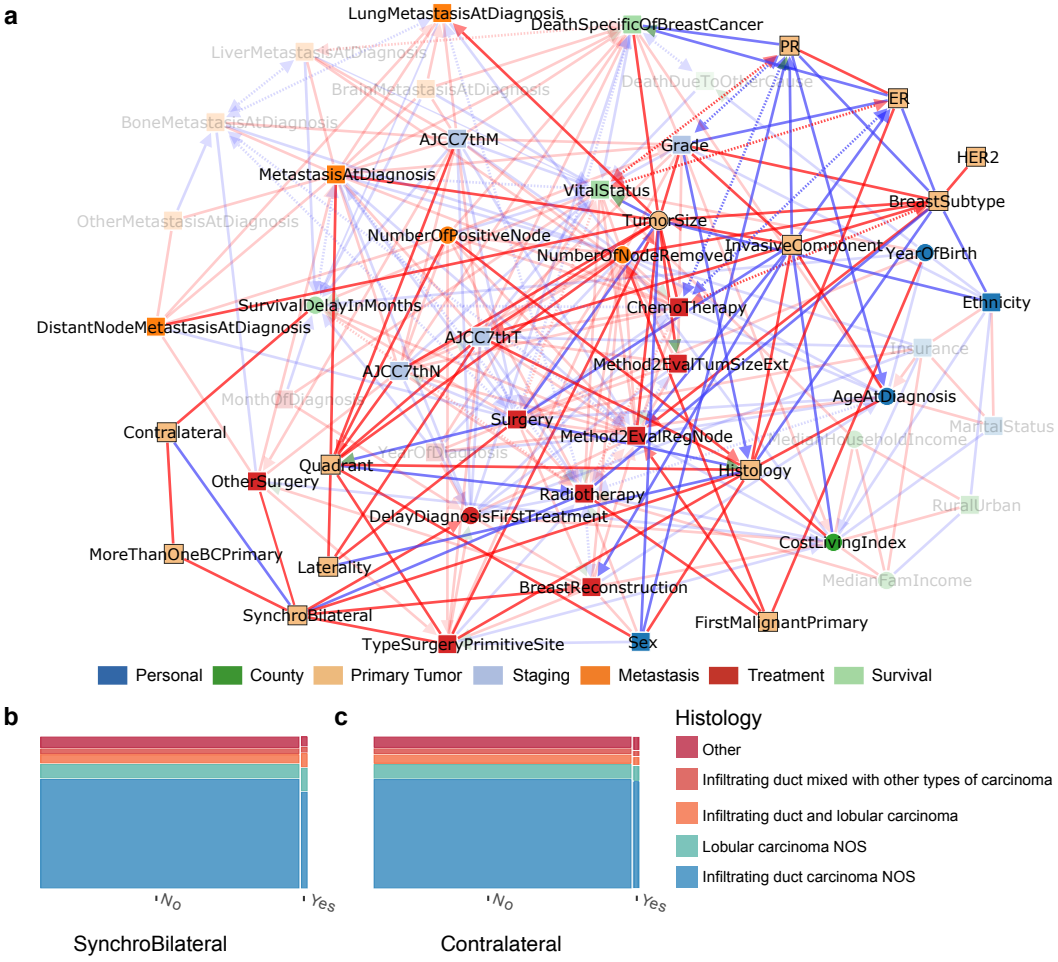


Fig. S8. Primary Tumor subnetwork inferred by iMIIC from SEER breast cancer dataset. **(a)** Subnetwork highlighting direct relations with primary tumor variables (Contralateral, MoreThanOneBCPrimary, SynchroBilateral, Laterality, Quadrant, Histology, FirstMalignantPrimary, TumorSize, InvasiveComponent, PR, ER, HER2, and BreastSubtype). **(b)** Joint distribution of Histology and Synchro Bilateral tumor. **(c)** Joint distribution of Histology and Contralateral tumor, see main text.

Reliable causal discovery based on mutual information supremum principle for finite datasets

Vincent Cabeli, Honghao Li, Marcel da Câmara Ribeiro-Dantas, Franck Simon, Hervé Isambert
Institut Curie, Université PSL, Sorbonne Université,
CNRS UMR168, 75005 Paris, France
first-name.last-name@curie.fr

Abstract

The recent method, MIIC (Multivariate Information-based Inductive Causation), combining constraint-based and information-theoretic frameworks, has been shown to significantly improve causal discovery from purely observational data. Yet, a substantial loss in precision has remained between skeleton and oriented graph predictions for small datasets. Here, we propose and implement a simple modification, named conservative MIIC, based on a general mutual information supremum principle regularized for finite datasets. In practice, conservative MIIC rectifies the negative values of regularized (conditional) mutual information used by MIIC to identify (conditional) independence between discrete, continuous or mixed-type variables. This modification is shown to greatly enhance the reliability of predicted orientations, for all sample sizes, with only a small sensitivity loss compared to MIIC original orientation rules. Conservative MIIC is especially interesting to improve the reliability of causal discovery for real-life observational data applications.

1 Background

Constraint-based structure learning methods can, in principle, discover causal relations in purely observational data (Pearl, 2009; Spirtes, Glymour, and Scheines, 2000). This is theoretically feasible up to some independence equivalence classes, as the orientations of certain edges may only be uncovered through perturbative data and remain undetermined if only observational data is available. Yet, regardless of this theoretical limitation, it has long been recognized (Ramsey, Spirtes, and Zhang, 2006; Colombo and Maathuis, 2014) that orientations predicted by constraint-based methods are often unreliable, which has largely limited, in practice, the application of constraint-based methods to uncover causal relations in real-life observational data.

This causal uncertainty originates from the extensive number of steps and conditions that constraint-based methods, such as the original IC (Pearl and Verma, 1991) and PC (Spirtes and Glymour, 1991) algorithms, have to meet before they can infer edge orientation. Indeed, they must first learn an undirected skeleton, by uncovering (conditional) independences between all pairs of variables, before inferring the orientation of v-structures and finally propagating these orientations to other undirected edges. This long chain of uncertain computational decisions leads to the accumulation of errors which ultimately limit the accuracy of the final orientation and propagation steps of constraint-based methods. As a result, edge orientations significantly reduce the precision (or positive predicted value) of inferred causal graphs compared to their undirected skeleton. In addition, constraint-based methods are known to suffer from much lower sensitivity or recall (*i.e.*, true positive rate) than precision scores, in general (Colombo and Maathuis, 2014; Li et al., 2019). This is related to the fact that separating sets used to remove edges in the (early) steps of constraint-based methods are frequently not consistent with the final skeleton and oriented graphs (Li et al., 2019). They correspond to

WHY-21 @ 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

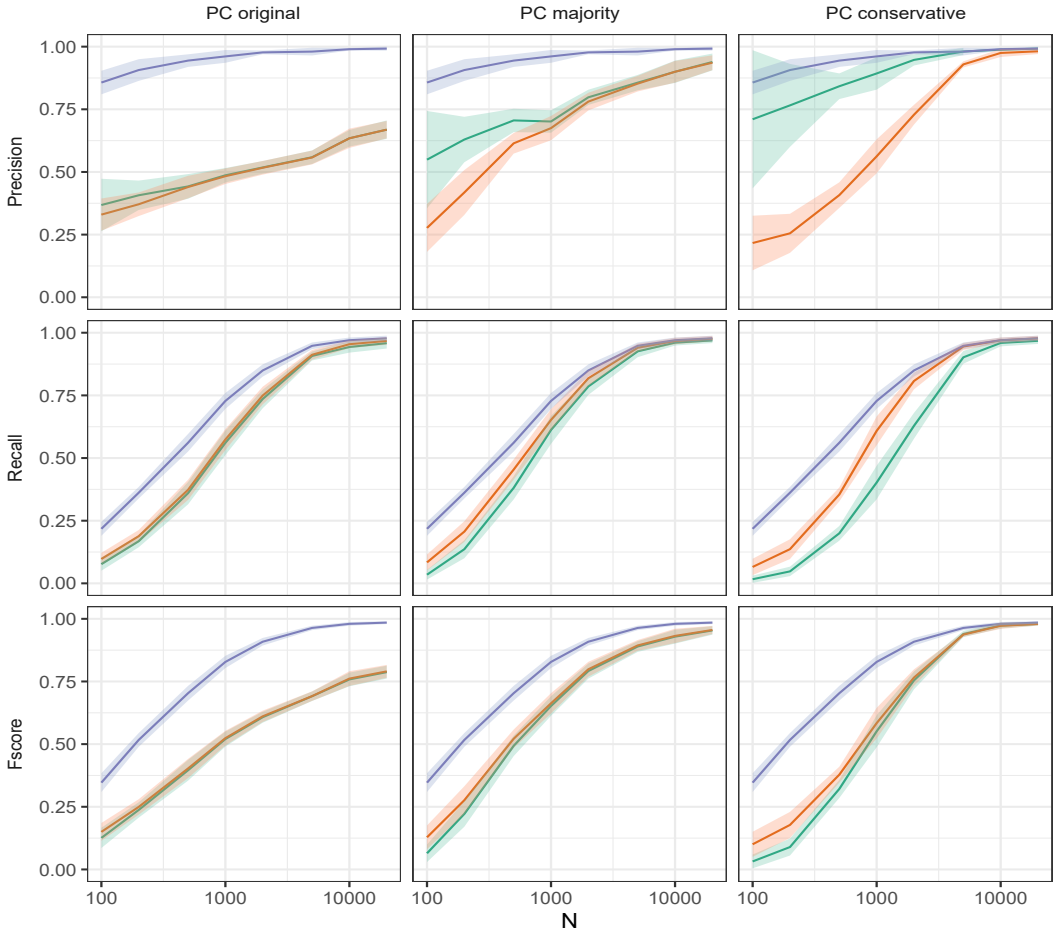


Figure 1: **PC original, majority and conservative orientation rules on discrete datasets**. Benchmark datasets are generated from random 100-node DAGs with average degree 3.8 and maximum degree 4 (See Data generation and benchmarks section for details). PC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green).

spurious conditional independences responsible for the large number of false negative edges and, therefore, low sensitivity of constraint-based methods.

While successive refinements of orientation rules, such as conservative rules (Ramsey, Spirtes, and Zhang, 2006) and majority rules (Colombo and Maathuis, 2014), have helped improve the average precision of orientations, they also lead to large precision variance and further aggravate the poor recall of edge orientations at small sample sizes. This is illustrated here for both discrete (Fig. 1) and continuous (Fig. 2) benchmark datasets generated by random Bayesian networks using the available codes from (Cabeli et al., 2020), see section on Data generation and benchmarks, below.

The recently developed method, MIIC, combining constraint-based and maximum likelihood frameworks, has been shown to significantly improve the situation by greatly reducing the imbalance between precision and recall, for all sample sizes (Verny et al., 2017; Cabeli et al., 2020). Compared to traditional constraint-based methods, MIIC also significantly reduces the precision gap between skeleton and oriented graphs for large enough datasets, as discussed below. However, a substantial loss in precision remains between skeleton and oriented graphs for smaller datasets.

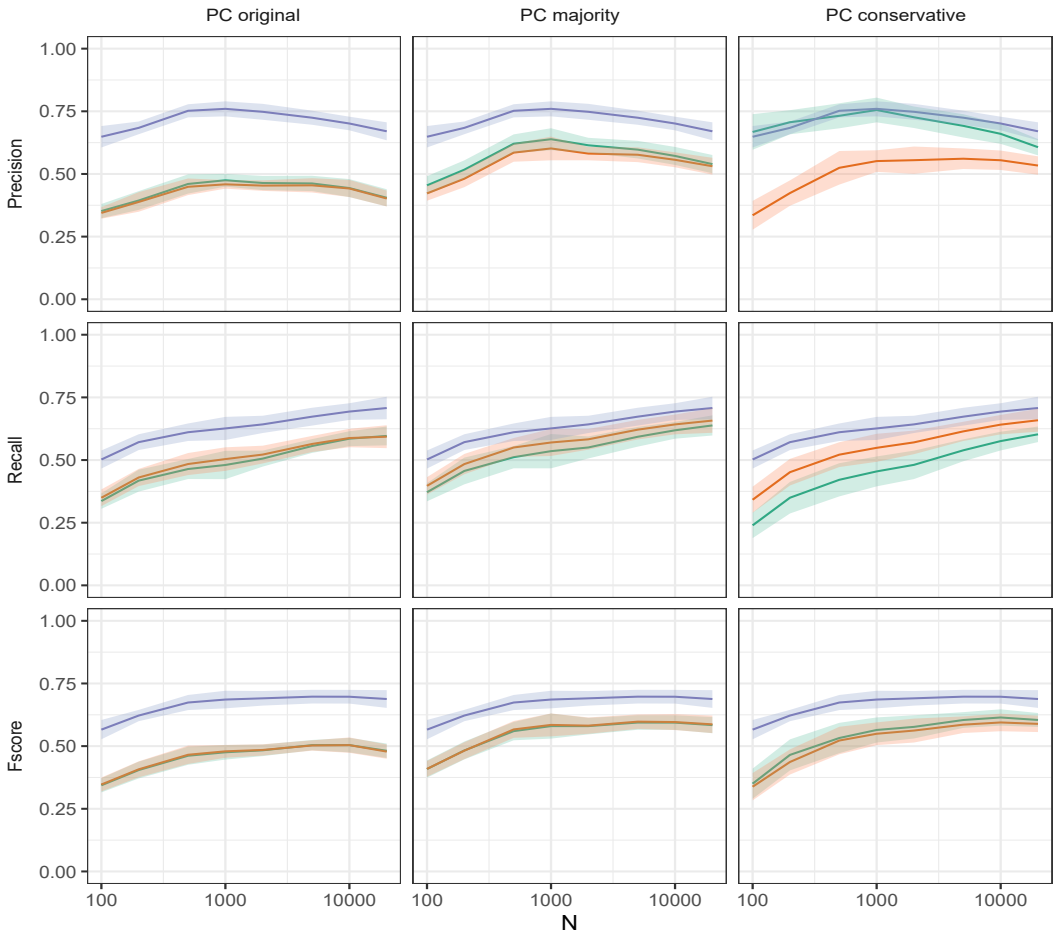


Figure 2: **PC original, majority and conservative orientation rules on continuous datasets**. Benchmark datasets are generated from random 100-node DAGs with average degree 3.8 and maximum degree 4 (See Data generation and benchmarks section for details). PC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green).

In this paper, we propose and implement a simple modification of MIIC algorithm, which is found to greatly improve the precision of predicted orientations even for relatively small datasets. It is achieved at the expense of a small loss of orientation recall but significantly enhances the reliability of predicted orientations for all sample sizes. This simple modification, referred to as conservative MIIC, is especially interesting, in practice, to improve the reliability of causal discovery for real-life observational data applications.

2 Results

2.1 MIIC outline

MIIC (Multivariate Information-based Inductive Causation) is a novel structure learning method (Verny et al., 2017; Cabeli et al., 2020) and online server (Sella et al., 2018), combining constraint-based and information-theoretic frameworks. Starting from a fully connected graph, MIIC iteratively removes dispensable edges, by uncovering significant information contributions from indirect paths based on the "3off2" scheme (Affeldt and Isambert, 2015; Affeldt, Verny, and Isambert,

2016). This amounts to progressively uncover the best supported conditional independencies, *i.e.* $I(X;Y|\{A_i\}_n) \simeq 0$, by iteratively "taking off" the most significant indirect contributions of *positive* conditional 3-point information, $I(X;Y;A_k|\{A_i\}_{k-1}) > 0$, from every 2-point (mutual) information, $I(X;Y)$, as,

$$I(X;Y|\{A_i\}_n) = I(X;Y) - I(X;Y;A_1) - I(X;Y;A_2|A_1) - \cdots - I(X;Y;A_n|\{A_i\}_{n-1}) \quad (1)$$

In practice, (conditional) independence is established by comparing mutual information (MI) or conditional mutual information (CMI) to a universal Normalized Maximum Likelihood (NML) complexity term, $k_N^{\mathrm{NML}}(X;Y|\{A_i\})/N$, computed over all datasets of the same size N and marginal distributions $p(X,\{A_i\})$ and $p(Y,\{A_i\})$ (Affeldt and Isambert, 2015). This can be seen as a NML-regularization of MI and CMI for datasets of finite sample size N as,

$$I'_N(X;Y|\{A_i\}) = I_N(X;Y|\{A_i\}) - \frac{1}{N}\, k_N^{\mathrm{NML}}(X;Y|\{A_i\}) \quad (2)$$

where $k_N^{\mathrm{NML}}(X;Y|\{A_i\})$ is computed iteratively in linear time (Kontkanen and Myllymäki, 2007; Roos et al., 2008) for increasing numbers of X and Y partitions, r_x and r_y, starting with $k_N^{\mathrm{NML}}(X;Y|\{A_i\}) = 0$ for $r_x = r_y = 1$ (Affeldt and Isambert, 2015; Cabeli et al., 2020).

Hence, (conditional) independence is established for $I'_N(X;Y|\{A_i\}) \leqslant 0$, whenever sufficient and significant indirect positive contributions could be iteratively collected in Eq. 1 to warrant the removal of the XY edge.

This leads to an undirected skeleton, which MIIC then (partially) orients based on the sign and amplitude of the NML-regularized conditional 3-point information terms (Affeldt and Isambert, 2015; Verny et al., 2017), corresponding to the difference between NML-regularized (C)MI terms.

$$I'_N(X;Y;Z|\{A_i\}) = I'_N(X;Y|\{A_i\}) - I'_N(X;Y|\{A_i\},Z) \quad (3)$$

In particular, negative NML-regularized conditional 3-point information terms, $I'_N(X;Y;Z|\{A_i\}) < 0$, correspond to the signature of causality in observational data (Affeldt and Isambert, 2015) and lead to the prediction of a v-structure, $X \to Z \leftarrow Y$, if $X - Z - Y$ is an unshielded triple in the skeleton (with $I'_N(X;Y|\{A_i\}) \leqslant 0$). By contrast, a positive NML-regularized conditional 3-point information term, $I'_N(X;Y;Z|\{A_i\}) > 0$, suggests to propagate the orientation of a previously directed edge $X \to Z - Y$ as $X \to Z \to Y$ (with $I'_N(X;Y|\{A_i\},Z) \leqslant 0$), to fulfill the assumptions of the underlying graphical model class.

2.2 MIIC performance on discrete data, allowing for negative NML-regularized MI & CMI

MIIC was originally developed for discrete variables only, for which MI and CMI are straightforward to compute. Compared to traditional constraint-based methods on discrete data, MIIC greatly reduces the imbalance between precision and recall, for all sample sizes, Fig. 3. MIIC also significantly reduces the precision gap between skeleton and oriented graphs, for large enough datasets. However, a substantial loss in precision remains between skeleton and oriented graphs, for small datasets, irrespective of the CPDAG or oriented-edge-only subgraph scores used for the comparison, Fig. 3.

These results illustrate the interest in integrating multivariate information criteria into constraint-based methods. However, for small datasets or datasets including variables with many discrete levels, NML complexities can easily out-weight MI and CMI terms for weakly dependent variables. As a result, MIIC tends to infer some v-structure orientations, $X \to Z \leftarrow Y$, for which both NML-regularized (C)MI terms in Eq. 3 are negative, *i.e.* $I'_N(X;Y|\{A_i\}) < I'_N(X;Y|\{A_i\},Z) < 0$, suggesting that Z could in fact be included in a separating set of the $\{X,Y\}$ pair, in contradiction with the inferred v-structure, $X \to Z \leftarrow Y$.

Note that such a v-structure would be excluded from the final graph in the frame of traditional constraint-based methods implementing conservative orientation rules, which check that Z is not included in any separating set of the $\{X,Y\}$ pair (Ramsey, Spirtes, and Zhang, 2006). Similarly, rectifying all negative NML-regularized (C)MI values into null values, as proposed and implemented in the present paper below, leads to a vanishing NML-regularized (conditional) 3-point information term in Eq. 3, *i.e.* $I'_N(X;Y;Z|\{A_i\}) = 0$, which precludes the orientation of the unshielded triple, $X - Z - Y$.

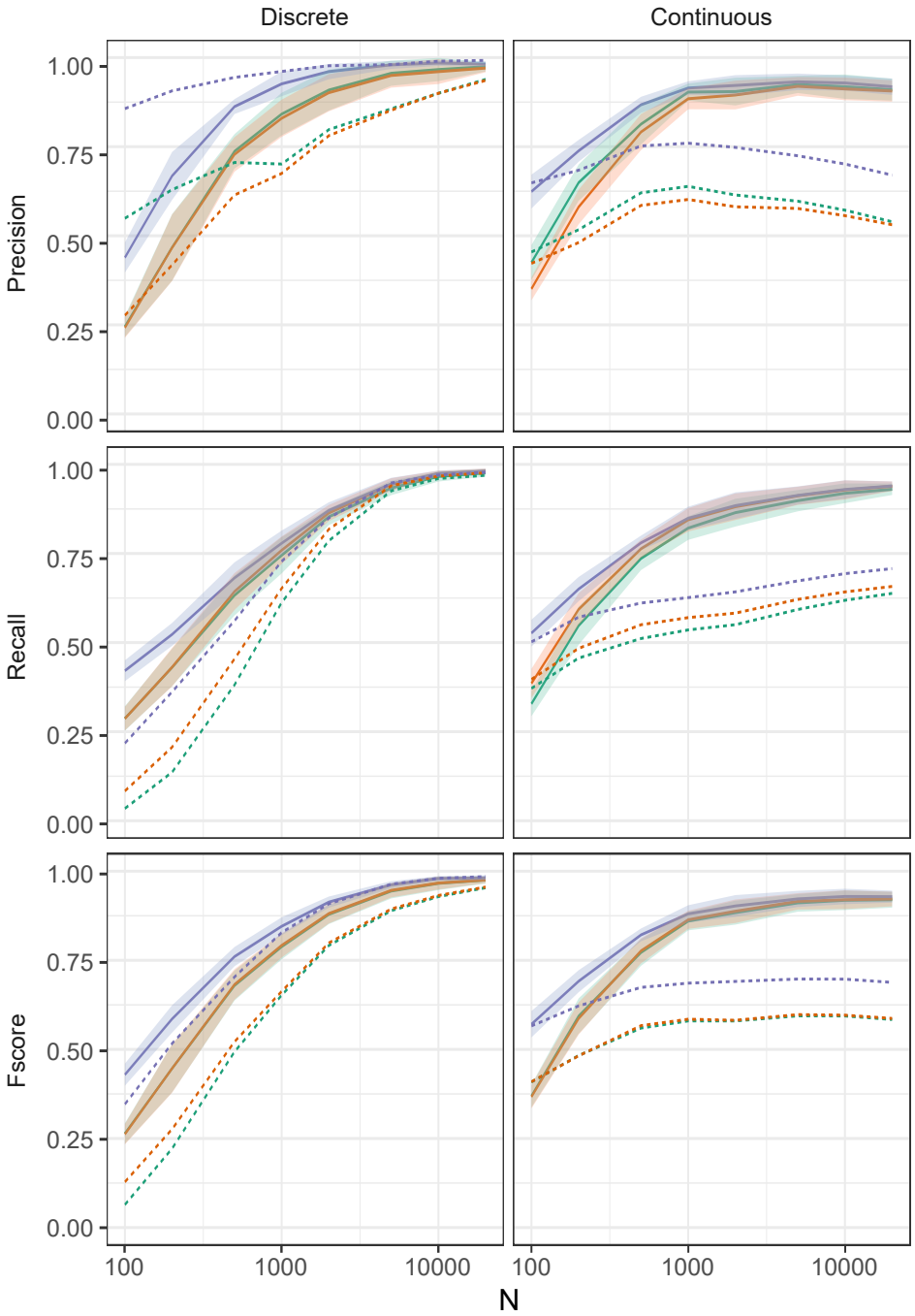


Figure 3: **Original MIIC with orientation rules allowing for negative NML-regularized MI & CMI on discrete data (left) and negative NML-regularized CMI on continuous data (right)**. Benchmark datasets are the same as in Figs. 1 & 2. MIIC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green). PC average scores for majority orientation rules are shown as dashed lines for comparison.

2.3 MIIC performance on continuous data, allowing for negative NML-regularized CMI

More recently MIIC was extended to handle continuous as well as mixed-type variables (either combination of discrete and continous variables or variables with both continuous and discrete ranges of values), for which MI & CMI are notoriously more difficult to estimate (Cabeli et al., 2020).

While distance-based k-nearest neighbor (kNN) estimates of MI and CMI are often used for continuous variables (Kraskov, Stögbauer, and Grassberger, 2004; Frenzel and Pompe, 2007), MIIC's MI and CMI estimates are instead computed through an approximate optimum discretization scheme, based on a general MI supremum principle (Cover and Thomas, 2006) regularized for finite datasets and using an efficient $\mathcal{O}(N^2)$ dynamic programming algorithm (Cabeli et al., 2020). This approach finds optimum partitions, $\mathcal{P}$ and $\mathcal{Q}$, specifying the number and positions of cut-points of each continuous variable, X and Y, to maximize the NML-regularized MI between them,

$$I'_N(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I'_N([X]_\mathcal{P};[Y]_\mathcal{Q}) \tag{4}$$

The NML regularization term, introduced in $I'_N([X]_\mathcal{P};[Y]_\mathcal{Q})$, is necessary for finite datasets and amounts to a model complexity cost, which eventually out-weights the information gain in refining bin partitions further, when there is not enough data to support such a refined model (Cabeli et al., 2020).

Such optimization-based estimates of MI are at par with alternative distance-based kNN approaches but have also the unique advantage of providing an effective independence test to identify independent continuous or mixed-type variables (Cabeli et al., 2020). This is achieved when partitioning X and Y into single bins maximizes the NML-regularized MI in Eq. 4, which vanishes exactly, in this case, with dramatic reductions in sampling error and variance (Cabeli et al., 2020). By contrast, kNN-MI estimates still need an actual independence test to decide whether some variables are effectively independent or not, as kNN MI estimates are never exactly null.

MIIC Precision, Recall and F-score on continuous data are comparable to those on discrete data, Fig. 3, and typically much better than the results obtained with traditional constraint-based methods, which, unlike MIIC, need to rely on independence tests, that are notoriously difficult for continuous data.

However, by contrast with discrete data, the remaining loss between skeleton and oriented graph precisions appears to differ between the CPDAG score and the oriented-edge-only subgraph score used for the comparison, Fig. 3. It indicates that the precision of the oriented-edge-only subgraph is slightly though significantly better than for the overall partially oriented graph, with a small concomitant loss of orientation recall, at small sample sizes, Fig. 3. This trend is due to the more stringent condition for v-structure orientation brought by the non-negative NML-regularized MI estimates obtained by MIIC for continuous variables. Yet, the optimum partitioning principle only applies to MI (Cover and Thomas, 2006), not CMI, which need to be estimated through the *difference* between optimum NML-regularized MI terms, as $I'_N(X;Y|U) = I'_N(Y;\{X,U\}) - I'_N(Y;U) = I'_N(X;\{Y,U\}) - I'_N(X;U)$ (Cabeli et al., 2020). As a result, the approximate NML-regularized CMI estimates between conditionally independent variables can sometime be negative and lead to v-structure orientations contradicting conditional independence, as discussed for discrete data above.

2.4 Improving MIIC causal discovery by rectifying negative NML-regularized MI & CMI

The general MI supremum principle (Cover and Thomas, 2006), regularized in Eq. 4 for finite datasets, is theoretically valid for any type of variables, not just continuous variables. In particular, it could be applied to small size datasets with discrete or categorical variables with many levels. It would result in the merging of rare levels to better estimate MI and CMI between weakly dependent discrete variables. Ultimately, MI estimates between independent discrete variables should lead to the merging of each variable into a single bin, thereby, resulting in NML-regulatized MI estimates to vanish exactly in this case, as already observed for continuous variables (Cabeli et al., 2020). As a result, optimum NML-regulatized MI should be non-negative as well as, by extension, NML-regulatized CMI, as shown now.

Theorem 1. *Optimum NML-regulatized MI and NML-regulatized CMI are non-negative.*

Proof. We first address optimum NML-regularized MI, noting that $I'_N(X;Y) \geqslant I'_N([X]_1;[Y]_1) = 0$, where $[X]_1$ and $[Y]_1$ are the X and Y variables partitioned into single bins, which leads to a vanishing

NML-regularized MI, as both MI and NML complexity cost are null, in this case, as $k_N^{\mathrm{NML}}(X;Y) = 0$ for $r_x = r_y = 1$ (Affeldt and Isambert, 2015).

Then, NML-regularized CMI is defined as the *difference* between optimum NML-regularized MI terms as, $I'_N(X;Y|U) = I'_N(Y;\{X,U\}) - I'_N(Y;U) = I'_N(X;\{Y,U\}) - I'_N(X;U)$. However, partitioning X and Y into a single bin leads to $I'_N(Y;\{X,U\}) \geqslant I'_N(Y;\{[X]_1,U\}) = I'_N(Y;U)$ and $I'_N(X;\{Y,U\}) \geqslant I'_N(X;\{[Y]_1,U\}) = I'_N(X;U)$ thus implying $I'_N(X;Y|U) \geqslant 0$ □

Following these considerations on the negativity of NML-regularized (C)MI with MIIC original orientation implementation, we propose a small modification, based on Theorem 1 and referred to as conservative MIIC, by analogy to the conservative orientation rules of traditional constraint-based methods (Ramsey, Spirtes, and Zhang, 2006), as noted above.

Proposition 2. *Conservative MIIC rectifies negative values of NML-regularized (C)MI, indicating (conditional) independence, to null values instead.*

The effects on this modification on discrete and continuous benchmark data are show in Fig. 4. While conservative MIIC hardly affects skeleton scores, it clearly has an impact on CPDAG and oriented-edge-only subgraph scores, which exhibit different trends relative to their original MIIC values.

CPDAG Precision, Recall and, hence, F-scores appear to be slightly lower under conservative MIIC (Fig. 4) than with original MIIC (Fig. 3), for discrete data. This illustrates the overall "better" orientation/non-orientation scores of the original MIIC against the theoretical CPDAG objective. Indeed, allowing for negative NML-regularized MI enables to infer weakly supported v-structures at small sample sizes. Besides, no significant difference is observed for CPDAG scores on continuous data, as original MIIC already enforces non-negative NML-regularized MI through optimization for continuous data (Cabeli et al., 2020), suggesting that enforcing also non-negative NML-regularized CMI with conservative MIIC has little impact on the reliability of CPDAG scores for continuous data, at least for the benchmarks tested here.

By contrast, conservative MIIC is found to greatly improve the precision of oriented-edge-only subgraphs, on discrete datasets, even for relatively small sample sizes, Fig. 4. This large increase in orientation precision is achieved at the expense of a relatively small loss of orientation recall. Hence, conservative MIIC significantly enhances the reliability and sensitivity of predicted orientations for all sample sizes, as compared to traditional constraint-based methods with conservative orientation rules, Fig. 4. For instance, conservative MIIC already reaches nearly 90% orientation precision with 25% orientation recall for $N \simeq 250$ (against about 80% orientation precision with only 5% orientation recall for conservative PC). While, by the time conservative PC reaches 90% orientation precision with 25% orientation recall for $N \simeq 700$, conservative MIIC achieves nearly 100% orientation precision with 50% orientation recall, Fig. 4. In addition, while original MIIC achieves a significantly better 65% orientation recall for $N \simeq 700$, Fig. 3, its orientation precision simultaneously drops to about 75%, which clearly impacts its reliability for causal discovery.

On continuous data, conservative MIIC also achieves a large increase in orientation precision, which becomes at par with skeleton precision, even for small datasets, and clearly much better than the corresponding scores obtained with traditional constraint-based methods for large datasets, Fig. 4. For instance, conservative MIIC reaches nearly 75% orientation precision with 50% orientation recall for $N \simeq 200$ (against about 70% orientation precision with 35% orientation recall for conservative PC). While, by the time conservative PC reaches 75% orientation precision with 45% orientation recall for $N \simeq 1,000$, conservative MIIC achieves more than 90% orientation precision with 80% orientation recall, Fig. 4.

3 Data generation and benchmarks

Datasets were simulated using structural equations models (SEMs) following the causal order of randomly generated DAGs. Continuous examples were constructed using linear and non-linear functions, and discrete datasets using unique state probabilities for each of the parents' combinations. The DAGs themselves were randomly drawn from the space of all possible 100 node DAGs (Melancon and Philippe, 2004) allowing for a maximum degree of 4 neighbors, resulting in an average degree of 3.8. Further details and dataset examples can be found in Cabeli et al. (2020).

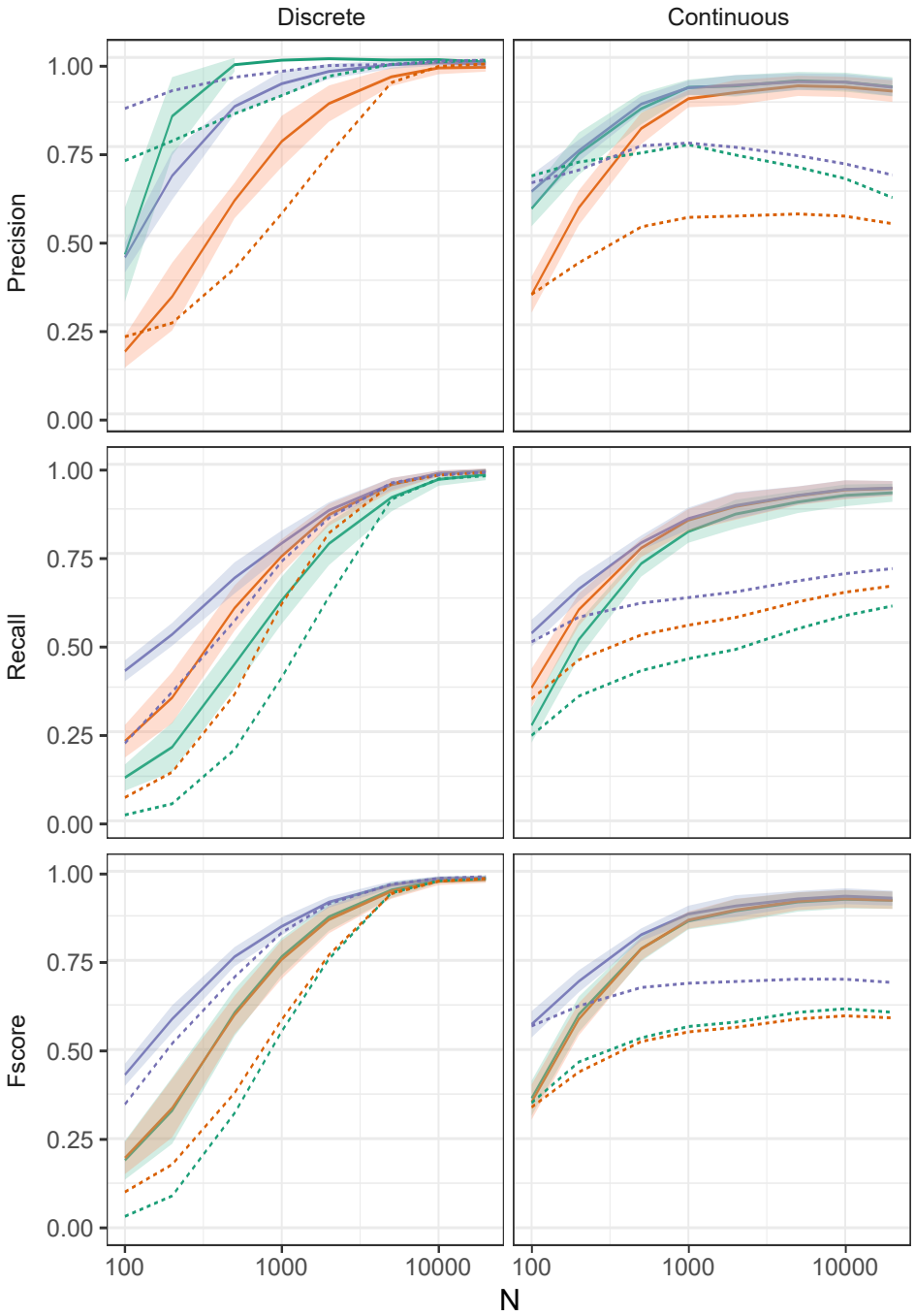


Figure 4: **Conservative MIIC with new orientation rules enforcing non-negative NML-regularized MI & CMI on discrete data (left) as well as continuous data (right)**. Benchmark datasets are the same as in Figs. 1 & 2. Conservative MIIC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green). PC average scores for conservative orientation rules are shown as dashed lines for comparison.

For evaluation purposes, network reconstruction was treated as a binary classification task and classical performance measures, Precision, Recall and F-score, were first used to evaluate skeleton reconstruction, based on the numbers of true *versus* false positive (TP *vs* FP) edges and true *versus* false negative (TN *vs* FN) edges, irrespective of their orientation.

Then, in order to evaluate edge orientations, we also define two orientation-dependent measures.

The first measure, referred to as the "CPDAG" score, aims to score the overall reconstruction with regards to the equivalence class of the true DAG. Edge types are used to redefine the orientation-dependent counts as, $TP' = TP - TP_{misorient}$ and $FP' = FP + TP_{misorient}$ with $TP_{misorient}$ corresponding to all true positive edges of the skeleton with a different orientation/non-orientation status as in the true CPDAG. The CPDAG precision, recall and F-score were then computed with the orientation-dependent TP' and FP'. In particular, the CPDAG score equivalently rates as "false positive" the erroneous orientation of an non-oriented edge in the CPDAG and the erroneous non-orientation of an oriented edge in the CPDAG. However, these errors are not equivalent from a causal discovery perspective.

The second measure, referred to as oriented-edge-only score, uses the same metrics but is restricted to the subgraphs of the CPDAG and the inferred graph containing oriented edges only. It is designed to specifically assess the method performance with regards to causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.

MIIC was run with default parameters for all settings on the latest version (available at `https://github.com/miicTeam/miic_R_package`), and PC with the `pcalg` package (Kalisch et al., 2012) using `bnlearn`'s (Scutari, 2010) mutual information test for discrete datasets and rank correlation for continuous ones. For PC, the α threshold for significance testing was tuned for each sample size N and network type to produce the best average between skeleton and "CPDAG" F-scores using a zeroth order optimization implemented in `dlib` (King, 2009).

4 Conclusion

Causal uncertainty and limited sensitivity of traditional constraint-based methods have so far hampered their dissemination for a wide range of possible causal discovery applications on real-life observational datasets. Hence, fulfilling the promise of causal discovery methods in the new data analysis area requires to improve their reliability as well as scalability.

We propose and implement, in this paper, a simple modification of the recent causal discovery method, MIIC, which greatly enhances the reliability of predicted orientations, for all sample sizes, with only a small sensitivity loss compared to MIIC original orientation rules. This conservative MIIC approach is especially interesting, in practice, to improve the reliability of cause-effect discovery for real-life observational data applications.

5 Acknowledgements

We would like to acknowledge the supports of the following funding agencies: Fondation ARC (VC), French Ministery of Higher Education, Research and Innovation (HL), EU IC-3i cofund PhD programme (MCRD), INSERM ITMO Cancer (FS, HI), CNRS and Institut Curie (HI).

References

Affeldt, S., and Isambert, H. 2015. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, 42–51.

Affeldt, S.; Verny, L.; and Isambert, H. 2016. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinformatics* 17(S2):12.

Cabeli, V.; Verny, L.; Sella, N.; Uguzzoni, G.; Verny, M.; and Isambert, H. 2020. Learning clinical networks from medical records based on information estimates in mixed-type data. *PLOS Computational Biology* 16(5):e1007866.

Colombo, D., and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* 15:3741–3782.

Cover, T. M., and Thomas, J. A. 2006. *Elements of Information Theory*. Wiley, 2nd edition.

Frenzel, S., and Pompe, B. 2007. Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.* 99:204101.

Kalisch, M.; Mächler, M.; Colombo, D.; Maathuis, M. H.; and Bühlmann, P. 2012. Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* 47(11):1–26.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10:1755–1758.

Kontkanen, P., and Myllymäki, P. 2007. A linear-time algorithm for computing the multinomial stochastic complexity. *Inf. Process. Lett.* 103(6):227–233.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Phys. Rev. E* 69:066138.

Li, H.; Cabeli, V.; Sella, N.; and Isambert, H. 2019. Constraint-based Causal Structure Learning with Consistent Separating Sets. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 14257–14266.

Melancon, G., and Philippe, F. 2004. Generating connected acyclic digraphs uniformly at random. *arXiv:cs/0403040*. arXiv: cs/0403040.

Pearl, J., and Verma, T. 1991. A theory of inferred causation. In *In Knowledge Representation and Reasoning: Proc. of the Second Int. Conf.* 441–452.

Pearl, J. 2009. *Causality: models, reasoning and inference*. Cambridge University Press, 2nd edition.

Ramsey, J.; Spirtes, P.; and Zhang, J. 2006. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, UAI, 401–408. Oregon, USA: AUAI Press.

Roos, T.; Silander, T.; Kontkanen, P.; and Myllymäki, P. 2008. Bayesian network structure learning using factorized nml universal models. In *Proc. 2008 Information Theory and Applications Workshop (ITA-2008)*. IEEE Press.

Scutari, M. 2010. Learning bayesian networks with the bnlearn r package.

Sella, N.; Verny, L.; Uguzzoni, G.; Affeldt, S.; and Isambert, H. 2018. Miic online: a web server to reconstruct causal or non-causal networks from non-perturbative data. *Bioinformatics* 34(13):2311–2313.

Spirtes, P., and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* 9:62–72.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, 2nd edition.

Verny, L.; Sella, N.; Affeldt, S.; Singh, P. P.; and Isambert, H. 2017. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.* 13(10):e1005662.

Posters et séminaires

Journées françaises de réseaux bayésiens 2023

Découverte causale sur de larges séries temporelles
Application à des séquences d'images d'écosystème tumoral

Franck Simon[1], Louise Dupuis[1], Vincent Cabeli[1], Tiziana Tocci[1], Nikita Lagrange[1], Hervé Isambert[1*]

1 CNRS UMR168, Institut Curie, Université PSL, Sorbonne Université, Paris, France
* corresponding authors: herve.isambert@curie.fr

Nous présentons ici la version temporelle de la méthode de découverte causale, MIIC[1,2], qui apprend des réseaux causaux pour un large éventail de données mixtes et incluant des variables latentes. MIIC a été appliqué à des données biologiques ou biomédicales, comme des données transcriptomiques et génomiques[1] ou des données de dossiers médicaux de patients[2]. Les séries temporelles, comme des images vidéo de cellules vivantes, contiennent cependant des informations sur la dynamique des variables d'intérêt, qui peut en principe faciliter la découverte de processus fonctionnels de cause à effet, en partant de l'hypothèse que les événements futurs ne peuvent pas causer des événements passés.

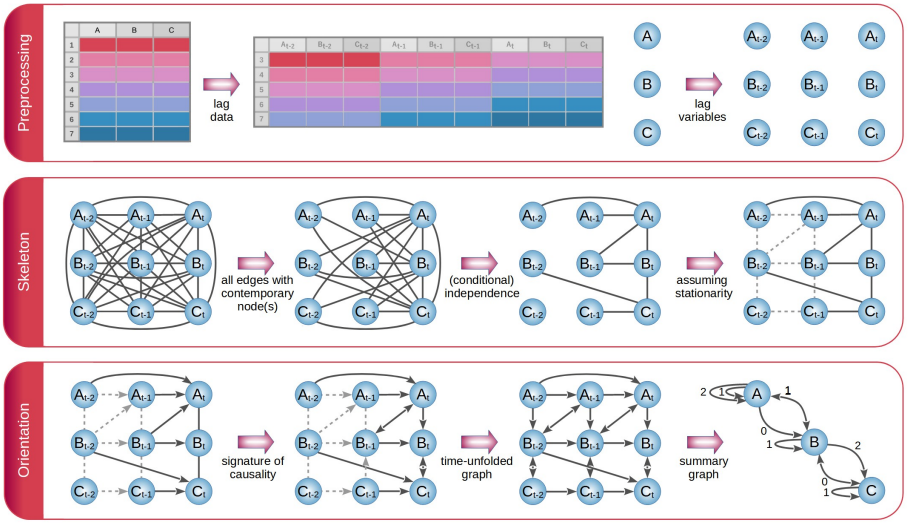


Figure 1. Schéma algorithmique de la version temporelle de MIIC[1,2] (Simon et al, en préparation).

Afin d'analyser des ensembles de données de séries chronologiques, la version temporelle de MIIC vise à apprendre un graphe déplié dans le temps, où chaque variable est représentée par une série de nœuds associés à sa valeur à différents moments relatifs, **Figure 1**. Un tel réseau déplié dans le temps est nécessaire pour tenir compte de la corrélation temporelle entre des échantillons successifs dans les données de séries chronologiques. En admettant que la dynamique peut être considérée comme stationnaire, le graphe recherché doit être invariant par translation dans le temps et une structure périodique peut lui être assignée *a priori*. De plus, la découverte causale peut être limitée à quelques pas de temps à partir du temps d'exécution, t, jusqu'à un maximum de $t - \tau_{max}$, pas de temps passés.

La version temporelle de MIIC a été utilisée pour effectuer l'analyse de séquences d'images d'un écosystème tumoral reconstitué *ex vivo* à l'aide de la technologie tumor-on-chip. Ces séquences proviennent d'une étude de preuve de concept[3] qui a déterminé les effets d'un médicament anticancéreux (les anticorps monoclonaux trastuzumab, nom de marque Herceptin, utilisés pour traiter des cancers du sein HER2+) sur un micro-environnement tumoral reconstitué comprenant des cellules cancéreuses, des cellules immunitaires, des fibroblastes associés au cancer (CAF) et des cellules endothéliales.

Les caractéristiques cellulaires telles que la géométrie cellulaire, la vitesse, la division, l'apoptose (*i.e.* mort cellulaire), les interactions transitoires et les contacts persistants entre cellules, ont été extraites des images brutes avec l'information de temps avec l'algorithme de segmentation et de tracking CellHunter[3]. Ces données ont été soumises en entrée de la version temporelle de MIIC pour inférer un réseau causal.

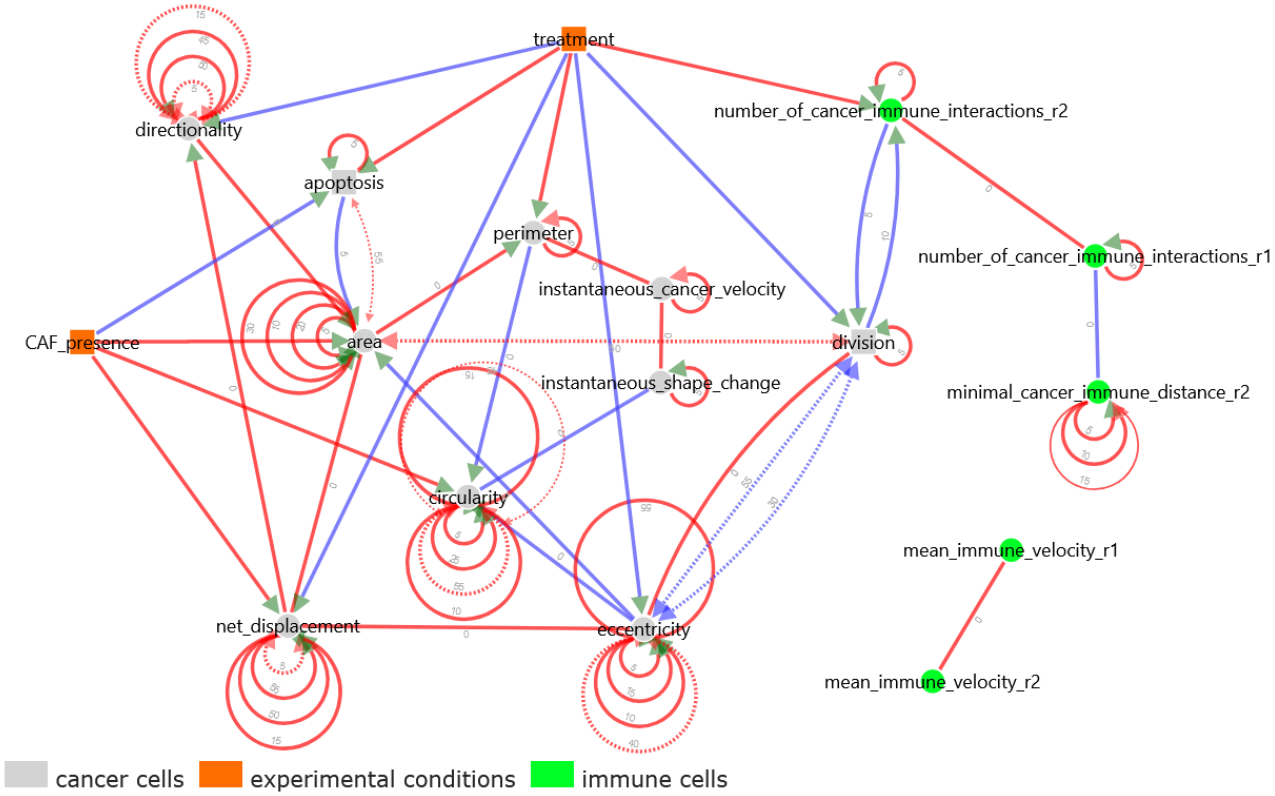


Figure 2. Réseau causal résumé obtenu à partir de données temporelles (50,000 images, dt=2 min) d'écosystème tumoral incluant des cellules cancéreuses et immunitaires dans 4 conditions: avec ou sans traitement et avec ou sans CAFs (Simon *et al*, en prép).

Le réseau causal temporel résumé obtenu, **Figure 2**, révèle de nouvelles découvertes biologiquement pertinentes en plus de confirmer les résultats connus d'études antérieures. En particulier, MIIC temporel découvre que les CAFs inhibent directement l'apoptose des cellules cancéreuses, indépendamment du traitement anticancéreux. MIIC temporel découvre également que le traitement réduit le périmètre des cellules cancéreuses et inhibe leur migration, ce qui n'a pas été rapporté jusqu'à présent non plus. De plus, MIIC temporel confirme les résultats connus d'études antérieures, il retrouve notamment que le traitement augmente l'apoptose des cellules cancéreuses et le nombre d'interactions entre cellules cancéreuses et immunitaires, et diminue le taux de division des cellules cancéreuses. De même, MIIC temporel retrouve l'effet que les CAFs stimulent la migration des cellules cancéreuses et augmentent leur aire et circularité.

Fait intéressant, MIIC temporel identifie également des effets multiples et éventuellement antagonistes avec des délais différents. Par exemple, MIIC temporel retrouve plusieurs relations antagonistes connues entre des caractéristiques morphodynamiques telles que la division cellulaire et l'excentricité. En effet, les phases tardives de la division cellulaire sont associées à une augmentation marquée de l'excentricité (arête rouge) mais la division cellulaire est précédée d'une nette diminution de l'excentricité, une à deux heures avant la cytocinèse (arêtes bleus), une fois que la décision de division a été prise (correspondant vraisemblablement à la cause latente) et que la cellule duplique en fait son matériel biologique (prophase).

Pour résumer, MIIC temporel permet de découvrir de nouvelles relations causales entre des caractéristiques cellulaires observées ou non et possiblement décalées dans le temps. Avec la disponibilité de données temporelles virtuellement illimitées dans de nombreux domaines, les méthodes de découverte causale sans hypothèse deviennent indispensables et nous pensons que MIIC temporel peut améliorer l'interprétation de ces données avec l'apport de la causalité.

Références

1. Verny L, Sella N, Affeldt S, Singh PP, Isambert H: Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput Biol* **13**(10):e1005662 (2017).

2. Cabeli V, Verny L, Sella N, Uguzzoni U, Verny M, Isambert H: Learning clinical networks from medical records based on information estimates in mixed-type data. *PLoS Comput Biol* **16**(5):e1007866 (2020).

3. Nguyen, M. *et al.* Dissecting Effects of Anti-cancer Drugs and Cancer-Associated Fibroblasts by On-Chip Reconstitution of Immunocompetent Tumor Microenvironments. *Cell Reports* **25**, 3884–3893.e3 (2018).

Imaging the Immune System

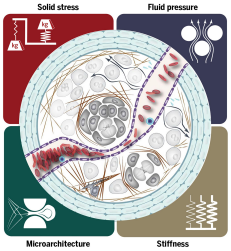
DCBIOL SEMINAR June 2023

Impact of macrophages on the tumoral extracellular matrix biophysical properties and consequences on the immune infiltrate

Zoé Fusilier, Lou Crestey, Mathilde Mathieu, Franck Simon, Valeria Manriquez, Perrine De Villemagne, Florence Piastra-Facon, Johanna Ivaska, Hervé Isambert, Christel Goudot, Ana-Maria Lennon-Duménil, Paolo Pierobon, Hélène Moreau

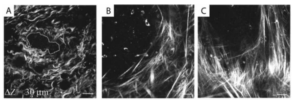
Scientific background & working hypothesis

What if tumor mechanics are as important as its chemistry ?



(Hadi T. Nia, Lance L. Munn, Rakesh K. Jain, 2020)

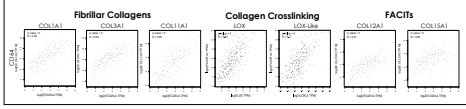
The extracellular matrix is profoundly remodeled during tumorigenesis



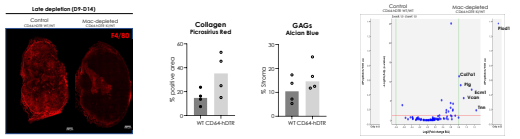
(Provenzano et al, 2006)

How are the biophysical properties of the ECM modified by the presence of macrophages during tumorigenesis ?
Does ECM remodeling by macrophages impact anti-tumoral immune response ?

CD64 expression correlates with the expression of structural ECM components in human colon adenocarcinoma (COAD database)

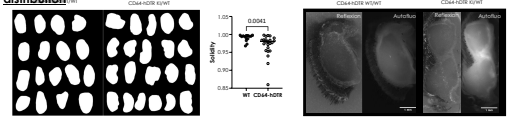


Macrophages-depleted MC38 subcutaneous tumors arbor altered ECM composition

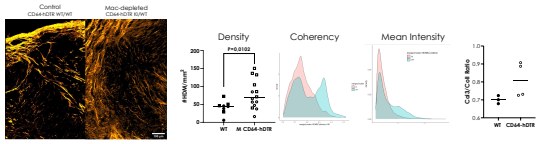


Results : Macrophage depletion modifies tumor morphology & ECM biophysical properties

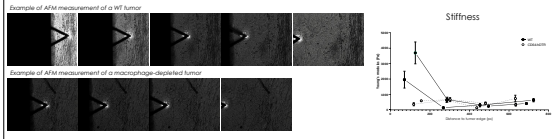
CD64hDTR late macrophage depletion modifies tumor morphology & ECM fiber spatial distribution



CD64hDTR late macrophage depletion modifies ECM fiber organization at the tumor edge



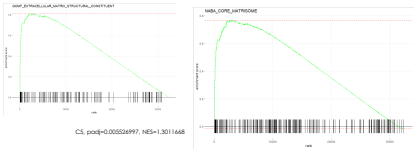
CD64hDTR late macrophage depletion disturbs stiffness gradient at the tumor edge



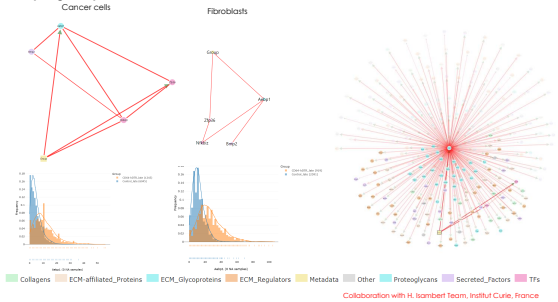
Collaboration with J. Ivaska Team, Turku Univ., Finland

Results : Macrophage depletion induces pro-ECM transcriptional programs in cancer cells and fibroblasts

Cancer cells overexpress ECM transcriptional programs in CD64-hDTR mice associated with ECM structure in macrophage-depleted tumors



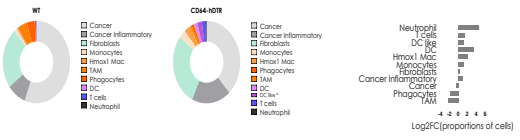
MIIC algorithm highlighted transcription factors potentially implicated in ECM remodeling in macrophage-depleted tumors



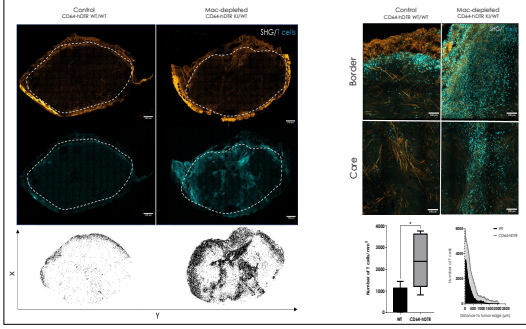
Collaboration with H. Isambert Team, Institut Curie, France

Results : Macrophage depletion modifies tumor immune infiltrate

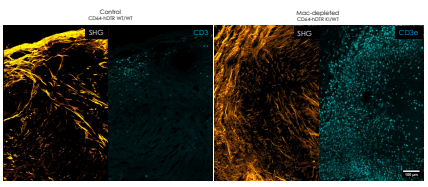
ScRNAseq: macrophage depletion increases immune cell infiltration



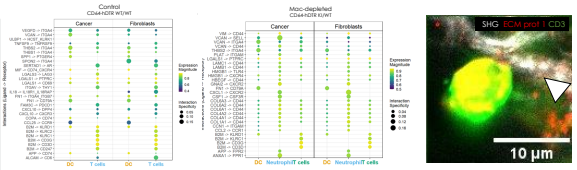
Macrophage depletion allows T cells and neutrophils (not shown here) to penetrate the tumor, in fiber- rich areas



Hypothesis #1 Mac-depletion modifies tumor ECM biophysical properties creating an optimal environment for T cell infiltration and function



Hypothesis #2 Mac-depletion increases immune cell-ECM interactions in the TME (in favor of the T cell anti-tumor immune response ?)



Inserm · institut Curie · FIRE PhD CRI · Université de Paris

CausalXtract: a flexible pipeline to extract causal effects from live-cell time-lapse imaging data

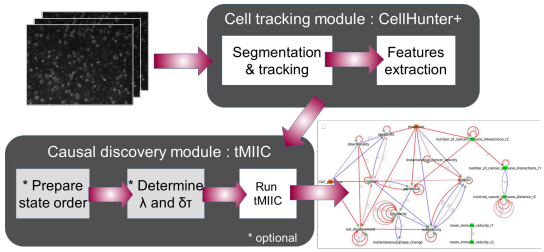
Franck Simon, Maria Colomba Comes, Tiziana Tocci, Louise Dupuis, Vincent Cabeli, Nikita Lagrange, Arianna Mencattini, Maria Carla Parrini, Eugenio Martinelli, Hervé Isambert

Causal discovery in time-lapse live-cell imaging data

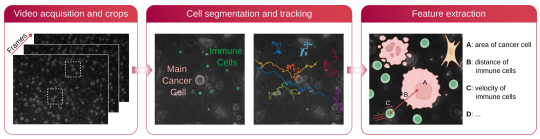
Live-cell imaging microscopy routinely produces a massive amount of time-lapse images of cellular systems. The individual cells can be segmented and tracked to extract morphodynamic features and cell-cell interactions under physiologically relevant conditions. However, it remains difficult to retrieve causal effect relationships from the features extracted due to the lack of methods and tools able to perform temporal causal discovery. We developed a novel flexible computational pipeline called CausalXtract that discovers causal and possibly time-lagged effects from live-cell time-lapse imaging data. It integrates an advanced cell image feature extraction tool, CellHunter+, with a reliable temporal causal discovery method, tMIIC. tMIIC stands for temporal Multivariate Information based Inductive Causation and aims to reconstruct robust temporal causal networks from large scale time-series datasets. Here we present the method developed and the network obtained for live-cell time-lapse images of tumor-on-chip cellular ecosystems under therapeutically relevant conditions. The inferred causal network gives precious insight on the *ex-vivo* tumor microenvironment, dissecting the effects of different experimental conditions on cancer cells' apoptosis and morphodynamics and on cancer-immune cell-cell interaction.

The method: CausalXtract

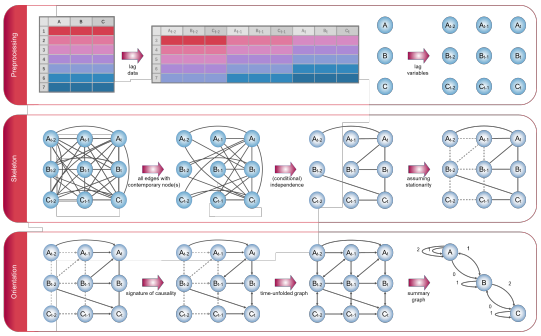
a. Complete pipeline to reconstruct causal graphs from images



b. CausalXtract's live–cell image feature extraction module (CellHunter+)



c. CausalXtract's temporal causal discovery module (tMIIC)

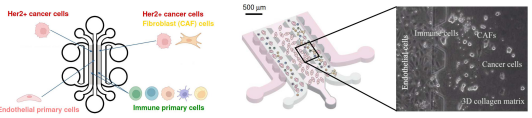


References

- Simon F, Comes MC, Tocci T, Dupuis L, Cabeli V, Lagrange N, Mencattini A, Parrini MC, Martinelli E, Isambert H, *CausalXtract: a flexible pipeline to extract causal effects from live-cell time-lapse imaging data*, submitted, 2023
- Nguyen M, De Ninno A, Mencattini A et al., Dissecting Effects of Anti-cancer Drugs and Cancer-Associated Fibroblasts by On-Chip Reconstitution of Immunocompetent Tumor Microenvironments, Cell Rep., 2018
- Verny L, Sella N, Affeldt S, Singh PP, Isambert H, *Learning causal networks with latent variables from multivariate information in genomic data*, PLoS Comput. Biol., 2017

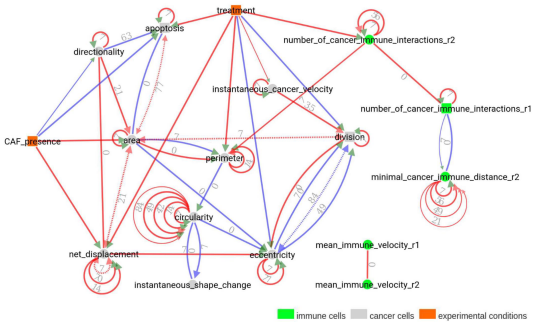
Example of application: tumor-on-chip

Tumor–on–chip preparation



Network inferred by CausalXtract

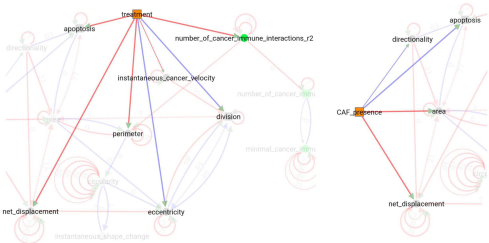
Summary causal network inferred by CausalXtract



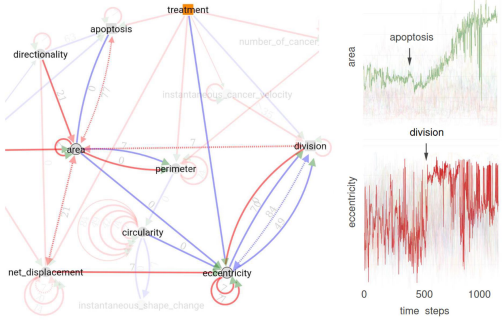
Subnetworks: treatment, CAF, time lagged effects

a. Treatment subnetwork

b. Fibroblasts subnetwork



c. Time lagged effects subnetwork



Created in BioRender.com