# Apprentissage de réseaux causaux avec variables latentes et applications à des contextes génomiques et cliniques

Louis Verny

Thèse présentée pour obtenir le grade de docteur
Université Pierre et Marie Curie

❧

UMR 168 - Physico-Chimie Curie
EDITE de Paris

**Discipline : Informatique**

---

# Apprentissage de réseaux causaux avec variables latentes et applications à des contextes génomiques et cliniques

---

PAR : **LOUIS VERNY**

**Sous la direction de** HERVÉ ISAMBERT, Institut Curie, Paris

MEMBRES DU JURY:

**Rapporteur** Denis THIEFFRY, Ecole Normale Supérieure, Paris

**Rapporteur** Étienne BIRMELÉ, Université Paris Descartes, Paris

**Examinateur** Pierre CHARBORD, Université Pierre et Marie Curie, Paris

**Examinateur** Leïla PERIÉ, Institut Curie, Paris

**Date de soutenance :** 04 décembre 2017

# REMERCIEMENTS

Je tiens tout d'abord à remercier Hervé Isambert, pour m'avoir donné ma chance au sein de son équipe et pour tout ce qu'il m'a apporté durant ces trois années.

Je souhaite également remercier sincèrement Messieurs Denis Thieffry et Étienne Birmelé, qui ont tous deux accepté d'être rapporteurs de mon travail de thèse. Je suis très content de bénéficier de leur expérience. Un grand merci aussi à Monsieur Pierre Charbord, ainsi qu'à Madame Leïla Perié, tous deux membres du jury, qui m'ont fait bénéficier de leur expérience et de leur connaissances au cours de discussions qui ont grandement profité à l'avancement de mes travaux.

Je souhaite remercier également Monsieur Fabien Tarissan, pour son aide précieuse lors de ma soutenance de mi-parcours.

Un grand merci à Séverine Affeldt, pour sa compagnie durant ma première année , pour sa disponibilité pendant sa thèse et après et pour son expérience de l'enseignement. Tout cela m'a beaucoup aidé.

Merci également à Nadir Sella, pour tous les moments passés à travailler ensemble en utilisant la technique désormais fameuse du "deux cerveaux pour un code", ce fut un réel plaisir et j'en garde un excellent souvenir.

Je voudrais dire aussi merci à Guido Uguzzoni, pour ses conseils sur la théorie mais surtout pour tous les moments de franche rigolade au bureau avec Nadir ainsi qu'à la boxe.

Un grand merci également au Dr. Fabien Reyal pour son enthousiasme, le temps qu'il m'a accordé et pour avoir permis cette collaboration qui m'a beaucoup appris. Je souhaite aussi remercier le Dr. Anne-Sophie Hamy-Petit également pour cet enthousiasme très communicatif, ainsi que pour sa patience et sa disponibilité. Vous avez été supers tous les deux.

Je tiens également à remercier les membres passés et présents du cercle très fermé

du Bureau 121AB : Démosthène, pour sa gentillesse, sa patience et ses conseils ; Marine, pour nos conversations animées et pour avoir été une oreille attentive ; Alicia, pour sa bonne humeur permanente et communicative ; Marvin, pour m'avoir forcé à stimuler mon sens de la répartie et pour sa bonne humeur ; Fazal, pour les travaux que nous avons réalisés ensemble et pour son entrain au quotidien et enfin, le dernier arrivé au sein de ce club privilégié Kevin, pour les discussions matinales et pour sa patience.

Un immense merci à Aurore, ma compagne, qui m'a soutenu tout au long de mon parcours, qui m'a beaucoup aidé à prendre les meilleures décisions pour mon avenir, et par dessus tout pour les délires et les fous rires au quotidien. Ça n'aurait pas été possible sans toi.

Un grand merci à mon Papa et à ma Maman, pour leur soutien sans faille et pour m'avoir si bien entouré, mais également pour leur avis professionnel et leur contribution au travail que je vais présenter.

Je remercie également du fond du cœur mes trois frangins, Maxime, Rémi et Thibaud, pour tout ce qu'ils m'apportent au quotidien et pour leur soutien.

Pour terminer, je souhaite dire un grand merci à mes potes de toujours : Cyril pour les pizzas du midi (et maintenant du soir) ; Jean-Nicolas pour les coups de fils sur le pouce et les déjeuners/dîners aménagés au milieu de nos agendas de ministres ; et Benjamin pour tous les supers moments qu'on passe ensemble (Bonne chance pour ton nouveau boulot !).

*A man's reach should exceed his grasp, or what's a heaven for ?*

Robert Browning

# INTRODUCTION

Le travail que j'ai effectué durant mes trois années de thèse dans l'équipe de Hervé Isambert concerne la reconstruction de modèles graphiques causaux sur des ensembles de données "réelles" finis, obtenus grâce à des expériences sur des systèmes biologiques ou *via* la mise en place d'essais cliniques selon le contexte. Ce manuscrit sera donc divisé en deux grands axes. Le premier axe concerne la mise au point de méthodes d'inférence de modèles graphiques et il est lui-même décomposée en deux parties : les méthodes tenant pas compte des effets des variables latentes d'une part, et les méthodes qui au contraire les prennent en compte d'autre part. Le second axe de mes travaux regroupe quant à lui l'analyse des résultats obtenus sur un ensemble d'applications concrètes dans des contextes génomiques et cliniques.

## 0.1 Reconstruction de modèles graphiques

L'apprentissage de réseaux causaux permet l'analyse graphique d'un jeu de données, en le caractérisant de façon visuelle et didactique. Un réseaux est ainsi constitué de deux catégories d'éléments, les variables (ou nœuds) et les liens (interactions, associations). Un modèle graphique causal permet de mettre en avant plusieurs caractéristiques du modèle sous-jacent.
Une interaction dite "directe" entre deux variables est représentée par un lien entre les deux nœuds correspondants sur le réseau : leur corrélation n'est alors pas due aux autres variables du réseau. En revanche, deux variables qui ne sont pas connectées directement sont soit totalement indépendantes (si aucun chemin indirect ne les relie), soit conditionnellement indépendante, dans le cas où d'autres variables servent d'intermédiaires. Pour finir, l'inférence de la **causalité** permet de savoir quelle variable influence l'autre lorsqu'elles sont reliées par une interaction directe.
Dans le cadre de la reconstruction de modèles graphiques, la causalité n'est pas définie comme une causalité **déterministe** ; qui implique que lorsque la cause est vraie, sa conséquence l'est également, sans incertitude possible ! Elle est définie comme une causalité plutôt **probabiliste**, c'est-à-dire qu'il sera plus probable d'observer la conséquence

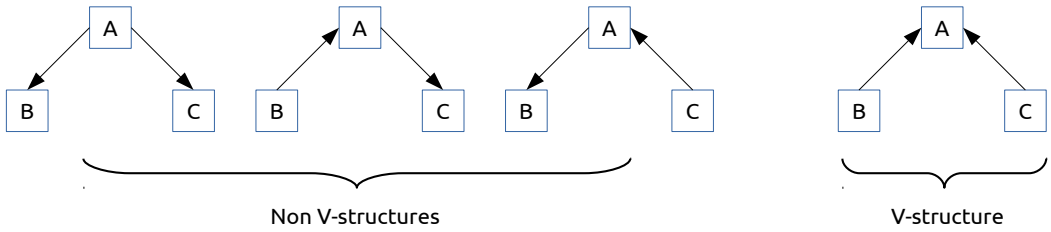



FIGURE 1 – **Différentes configurations de la causalité pour un triplet ouvert** – Sur des données d'observation seule la v-structure peut être différenciée lors de l'inférence de causalité, car les autres motifs reflètent tous la même indépendance conditionnelle, $C \perp\!\!\!\perp B|A$ (voir chapitre 1.1.2).

lorsque la cause a lieu (il est par exemple plus probable d'être victime du cancer du poumon lorsque l'on fume). De fait, l'une des façons de connaître la causalité serait de suivre un système au cours du temps, puisqu'il semble logique que la cause ait lieu avant sa conséquence. L'inférence de causalité dans le cadre de données d'observations non temporelles, bien que peu intuitive, n'est cependant pas un concept nouveau ; il a en effet déjà été mis en évidence par Pearl (1988) [30] auparavant, *via* la définition d'un motif particulier : les **v-structures**.
En effet selon Pearl, dans le cadre de données d'observation, une causalité peut être déduite lorsque l'on considère 3 variables. Si ces 3 variables ne sont reliées que par **2 liens** (formant alors un **triplet ouvert**), il est possible de détecter un motif causal particulier, appelé **v-structure**. C'est donc à partir de la détection de ce motif que l'on est capable d'inférer la causalité sur des données d'observation.

## 0.2 L'algorithme `MIIC`

Lorsque je suis arrivé dans l'équipe de Hervé Isambert, l'implémentation d'une nouvelle méthode de reconstruction de réseaux, baptisée `3off2`, était en train de voir le jour. Cet algorithme a pour objectif de tester chaque paire de variables, pour déterminer si l'information qui les relie peut être expliquée par l'action d'une ou de plusieurs variables intermédiaires en utilisant la théorie de l'information. Lorsque l'ensemble de ces tests est réalisé, on obtient un réseau non orienté, sans causalité, appelé **squelette**. Après obtention de ce squelette, l'algorithme recherche les **v-structures** afin d'orienter le réseau

(voir chapitre 4). Le fonctionnement de cet algorithme repose cependant sur deux hypothèses notables. Premièrement, les échantillons observés sont entièrement indépendants les uns des autres (pas de données issues de séries temporelles, donc). Deuxièmement, on suppose que le système observé est complet, c'est-à-dire qu'aucune variable susceptible d'avoir une influence quelconque n'est **cachée**.
Cette deuxième hypothèse en particulier peut amener l'algorithme à faire certaines erreurs lorsqu'elle n'est pas vérifiée, comme démontré dans le chapitre 5.2. Or ce cas n'est pas rare lorsque l'on s'intéresse à de véritables données expérimentales ; il est en effet parfois impossible, dans ce cas, d'observer l'ensemble des variables pouvant influer le système. C'est dans le but de résoudre cette problématique que l'algorithme `MIIC`, sur lequel j'ai travaillé durant ces trois années, se détache de cette exigence tout en reprenant le schéma de fonctionnement de `3off2`. Nous verrons que `MIIC` présente globalement de meilleurs résultats que la dernière variante de l'algorithme FCI (RFCI), méthode référence sur ce sujet. `MIIC` adresse également la problématique de l'indépendance des échantillons, puisqu'il est capable de détecter certaines corrélations entre les échantillons et de les prendre en compte ensuite grâce au calcul d'un **nombre efficace d'échantillons**. Cette modification permet d'obtenir des résultats satisfaisants, y compris dans le cadre de l'analyse de données de séries temporelles (voir section 7.1.3).

## 0.3 Analyse des modèles inférés dans les contextes biologiques et cliniques

A ce titre, l'exploration du domaine des réseaux de régulations transcriptionnelles m'a semblé très pertinente, puisqu'elle est l'un des sujets biologiques abordant le plus les questions d'inférence et de découverte de réseaux, à la fois dans les domaines de l'expérimentation et dans les domaines plus théoriques de la biologie computationnelle. Je me suis donc tout d'abord intéressé à la régulation transcriptionnelle de l'hématopoïèse adulte puisque ce processus, très étudié, s'est avéré idéal pour évaluer les performances de `3off2` et les comparer aux techniques déjà existantes. Cette analyse a été publiée dans la revue *BMC Bioinformatics* [2].
Une fois qu'il a été démontré que l'algorithme est capable d'inférer des interactions cohérentes sur des jeux de données expérimentales, j'ai souhaité rechercher le ou les avantages que pourraient apporter l'utilisation de la reconstruction de réseaux causaux sur de

tels jeux de données. C'est donc dans le but de répondre à cette problématique que j'ai analysé des données issues de deux domaines précis, la biologie d'une part (application sur l'hématopoïèse embryonnaire) et la médecine d'autre part, avec les applications sur les démences des personnes âgées et sur le cancer du sein.

### 0.3.1 Hématopoïèse embryonnaire

L'objectif de cette application a été de reconstruire, mais également de vérifier, le modèle de différentiation des premières cellules sanguines de l'embryon. Grâce à l'utilisation de méthodes complémentaires (analyse en composante principale (ACP) et clustering), nous avons été capables de retrouver et de caractériser sur le plan **transcriptionnel** les trois populations cellulaires définies par le modèle. De plus, nous apportons également certaines nuances intéressantes qui pourront éventuellement être explorées expérimentalement (cf section 8.3).

### 0.3.2 Données cliniques

Nous démontrerons également les avantages de l'application de `MIIC` pour des problématiques de santé publique telles que les démences chez les personnes âgées, ou le cancer du sein. Nous avons pu étudier le sujet des démences chez les personnes âgées grâce à une collaboration avec le service de gériatrie de la Pitié-Salpétrière, dirigé par le Pr Marc Verny. En analysant les données cliniques mises à notre disposition, nous avons pu reconstruire un modèle graphique robuste. Celui-ci souligne un grand nombre d'interactions déjà connues des spécialistes du domaine et met également en évidence une association non publiée dans ce contexte jusqu'à cette année.
Le travail sur le sujet du cancer du sein a quant à lui été réalisé dans le cadre d'une collaboration avec l'équipe du Residual Tumor and Response to Treatment Laboratory (Dr. Fabien Reyal et Dr. Anne-Sophie Hamy-Petit). Le réseau reconstruit sur les données de la cohorte *Neorep* permet également de visualiser des interactions communes, évidentes ou connues depuis longtemps, ainsi que les dernières associations en passe d'être publiées, découvertes par les spécialistes travaillant sur ces données. On découvre également une association surprenante, non publiée, qui pourra éventuellement ouvrir d'autres pistes de recherche sur ces données.

# TABLE DES MATIÈRES

PREMIÈRE PARTIE

# Apprentissage de modèles graphiques sans variable latente

CHAPITRE 1

# MODÈLES GRAPHIQUES : ASPECTS THÉORIQUES ET NOTATION

Ce chapitre a pour objectif de décrire les notions de la théorie des réseaux utilisées dans ce manuscrit.

## 1.1 Directed Acyclic Graphs (DAG)

### 1.1.1 Notions de base et terminologie

Un modèle graphique $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ est une "carte" des interactions entre un ensemble de variables aléatoires, représenté par $\mathbf{V} = \{X_1, X_2, ...., X_q\}$. Chaque association (ou lien) présent dans **E** signifie donc qu'il existe une interaction entre la paire de variable en question. Si **E** contient tous les liens possibles, le graphe est dit complet. Ces liens peuvent être de différents types :

- **dirigés** ($\rightarrow$) : indique le sens de la causalité. Ici une variable est la cause de l'autre, et influence donc son comportement (c'est-à-dire sa distribution de probabilités).
- **non dirigés** ($-$) : ceci indique soit que la causalité ne peut être découverte avec les données dont on dispose, soit en fonction du contexte, que l'association est due à une variable de sélection. Cette notion est abordée plus loin dans ce manuscrit.
- **bi-dirigés** ($\leftrightarrow$) : une double orientation représente l'existence d'une cause commune aux deux variables reliées, absente du jeu d'observations. La problématique posée par de telles variables est également abordée plus loin.

Lorsque tous les liens sont dirigés, le graphe est dit **dirigé**. Un modèle comprenant à la fois des liens dirigés et non-dirigés est dit **partiellement orienté**. Un graphe ne comprenant aucune orientation est appelé **squelette**.

A chaque nœud est donc associé un ensemble **d'adjacences**, correspondant à l'ensemble des liens qui sont directement connectés à un nœud $X_i$ dans le réseau $\mathcal{G}$. De plus, s'il

existe une relation telle que $X_i \rightarrow X_j$, $X_i$ est dit **parent** de $X_j$. Une chaîne de nœuds connectés est appelée **chemin**. Si la totalité des liens de ce chemin sont orientés dans la même direction, il est dit **dirigé**. Enfin, dans le cas où le chemin est dirigé de $X_i$ vers $X_j$ et qu'il existe un lien dirigé $X_j \rightarrow X_i$, le chemin forme un **cycle dirigé**. Si un graphe (partiellement) orienté ne contient pas de cycle dirigé, il est appelé **(P)DAG** (pour **Partially Directed Acyclic Graph**).

Dans ce type de modèle graphique, certains triplets de nœuds sont dits **ouverts**, ce qui signifie que seules deux paires sont adjacentes (seuls deux des trois liens possibles sont inclus dans **E**). Dans le cas où ce triplet est orienté de la façon suivante : $X_i \rightarrow X_k \leftarrow X_j$, il est appelé **v-structure** ; tandis que les trois autres configurations d'orientations possibles ($X_i \rightarrow X_k \rightarrow X_j$, $X_i \leftarrow X_k \rightarrow X_j$, $X_i \leftarrow X_k \leftarrow X_j$) sont des **non-v-structures**.

Comme détaillé dans les sections suivantes, ces trois configurations rendent compte du même schéma d'**indépendance conditionnelle** (ici $X_i \perp\!\!\!\perp X_j | X_k$), ce qui signifie qu'elle ne peuvent statistiquement pas être différenciées les unes des autres en se basant sur des données d'observations.

### 1.1.2 Interprétation des modèles graphiques acycliques et dirigés (DAG)

L'ensemble des indépendances conditionnelles des variables du graphe $\mathcal{G}$, contenues dans l'ensemble des variables **V**, peut être déduit de celui-ci en appliquant un critère formulé par Pearl en 2000 ([29]), appelé **critère de d-separation**. Il stipule que si deux variables (qu'on note $X_i$, $X_j$) ne sont pas adjacentes dans le DAG $\mathcal{G}$, elles sont **d-séparées** par **S**, un sous-ensemble des variables restantes de $\mathcal{G}$. Ceci implique que les variables $X_i$ et $X_j$ sont indépendantes, lorsque l'on les conditionne sur **S** ($X_i \perp\!\!\!\perp X_j | S$).

La définition de ce critère est liée à la notion d'**interruption du flux d'information** entre deux variables d'un modèle graphique donné. Prenons le cas de trois variables $X_i,\ X_j,\ X_k$ reliées par un triplet ouvert. Si ce triplet est une non-v-structure, les deux nœuds situées à ses extrémités sont **marginalement dépendants**. En revanche, lorsque l'on connaît la valeur de $X_k$ (la valeur est fixée), la connaissance de $X_i$ n'a pas d'impact supplémentaire sur notre connaissance de $X_j$ ; autrement dit les variables $X_i$ et $X_j$ deviennent indépendantes lorsque $X_k$ est connue : $X_k$ peut donc bloquer le chemin entre $X_i$ et $X_j$.

Dans le cas où ce triplet est une v-structure ($X_i \rightarrow X_k \leftarrow X_j$), les deux variables à l'extrémité du triplet sont *a priori* marginalement indépendantes ; mais deviennent dépendantes

si on les conditionne sur $X_k$.
Pour illustrer ceci, considérons par exemple les trois variables suivantes : $alarme$, $tremblement\ de\ terre$, $cambriolage$. Il est raisonnable de penser que ces variables sont reliées par une v-structure telle que $cambriolage \rightarrow alarm \leftarrow tremblement\ de\ terre$. Il est également raisonnable de dire qu'il n'existe *a priori* pas d'association entre un tremblement de terre et le fait de se faire cambrioler. Supposons maintenant que l'on sait que l'alarme s'est déclenchée, mais également qu'il y a eu un tremblement de terre au même moment : la probabilité que la maison ait été cambriolée devient très faible, puisque c'est vraisemblablement le tremblement de terre qui a déclenché l'alarme. Cet effet est appelé "explain away effect" ([1]).

### 1.1.3 Équivalence de Markov

Comme introduit précédemment dans la section 1.1.1, les 3 non-v-structures sont équivalentes en terme d'indépendances conditionnelles ; ce qui implique que différents DAG causaux peuvent également être équivalents dans ce sens ; ils sont alors dits **équivalents markoviens**. Formellement, deux graphes sont dit équivalents markoviens s'ils possèdent le même squelette et les mêmes v-structures. Le regroupement de l'ensemble des graphes équivalents est appelé **classe d'équivalence de Markov**.
Les classes d'équivalence de Markov sont représentées au moyen d'un **Graphe Acyclique Dirigé Partiellement Complété** (**CPDAG**), possédant les propriétés suivantes :

- tous les liens du CPDAG existent dans tous les DAG de la classe d'équivalence correspondante
- tous les liens non dirigés d'un CPDAG sont orientés de façon opposée dans au moins deux DAG appartenant à la classe d'équivalence correspondante.

Si on suppose que les hypothèses de **fidélité** (**faithfulness** en anglais) et de **suffisance causale** (**causal sufficiency**) sont réalisées (c'est-à-dire l'absence de variables de sélection non mesurées et de causes communes non mesurées, voir section 5.2), un CPDAG peut être reconstruit à partir des indépendances conditionnelles entre les variables aléatoires observées et utilisé pour représenter la classe d'équivalence des DAGs correspondant au modèle graphique causal sous-jacent.

C'est dans le contexte où les deux hypothèses précédemment mentionnées sont supposées vraies que sont définies les méthodes dites **bayésiennes**, aussi appelées **search and score**, présentées dans le chapitre 2, ainsi que la méthode `3off2`, mise au point dans l'équipe et publiée dans *BMC Bioinformatics* en 2016 [2] (voir chapitre 4).

CHAPITRE 2

# DESCRIPTION DES MÉTHODES "BAYÉSIENNES" : SEARCH AND SCORE

Le principe des méthodes bayésiennes est de parcourir l'espace des DAG correspondant à l'ensemble des variables observées grâce à des procédures appelées **search and score**. Une fonction de score est utilisée pour estimer la vraisemblance de chaque modèle graphique considéré par rapport aux données utilisées.

Plusieurs techniques sont utilisées dans ce but (Monte Carlo Markov Chain [20], Hill-Climbing), mais nous ne décrirons dans ce manuscrit que la méthode du Hill-Climbing.

## 2.1 Hill-Climbing

Étant donné que l'espace des configurations de graphes croît de façon super exponentielle en fonction du nombre de variables considérées, il est très rapidement impossible de le parcourir dans son ensemble. Cette problématique est généralement contournée par l'emploi de méthodes dites heuristiques, permettant la recherche d'une configuration correspondant à un maximum de la fonction de score pour les configurations testées, sans garantie qu'il s'agisse du maximum global. La méthode du Hill-Climbing se déroule de la façon suivante.

Premièrement, on considère un DAG (complètement orienté) au hasard et on calcule le score qui lui est associé ; puis on y insère un changement aléatoire (changement d'orientation d'un lien existant, ajout ou suppression d'un lien). Le score est recalculé sur la nouvelle configuration qui est soit conservée (si son score est meilleur que celui de la configuration précédente), soit retiré (si le score est inférieur à celui de la configuration précédente). Ces changements de configurations sont répétés jusqu'à ce qu'aucun changement n'améliore le score : le graphe correspondant est alors conservé.

L'utilisateur définit l'hyper paramètre $nbr\_restart$, qui représente le nombre de fois que cette procédure doit être réitérée, pour éviter les maximums locaux éventuels. Le graphe possédant le meilleur score parmi tous ceux qui ont été conservés est alors considéré comme le DAG le plus proche du DAG ayant généré les données.

## 2.2 Fonctions de score

Les 3 fonctions de score suivantes sont parmi les plus fréquemment utilisées dans ce cadre, et ce sont celles dont nous nous sommes servis pour mesurer les performances de cette classe de méthodes :

- Akaike information criterion (AIC) : proposé en 1974 par Akaike, ce critère basé sur la théorie de l'information quantifie l'information perdue lorsqu'on utilise un modèle $m$ pour approximer la réalité, en définissant également un terme pénalisant les modèles en fonction de leur nombre de paramètres. Sa définition formelle est la suivante[3] :

$$AIC = 2k - ln(\hat{L}) \tag{2.1}$$

  où $k$ représente le nombre de paramètres du modèle à estimer ; le terme $2k$ sert donc à pénaliser la complexité du modèle (évitant ainsi l'overfitting) et $\hat{L}$ représente la valeur du maximum de vraisemblance de ce modèle, c'est-à-dire :

$$\hat{L} = P(x|\hat{\theta}, M) \tag{2.2}$$

  où $\hat{\theta}$ correspond aux paramètres du modèle maximisant la fonction de vraisemblance. Pour être retenu, le modèle doit donc réaliser un compromis entre la fonction de vraisemblance (qui doit être la plus importante possible), et sa complexité (qui doit être minimum) : on cherche à minimiser le score AIC et donc la perte d'information vis-à-vis du phénomène que l'on cherche à modéliser.
- le Bayesian Information Criterion (BIC) : il est dérivé du score AIC, avec une modification de la fonction de pénalité qui dépend ici également de $n$, le nombre d'échantillons observés[37] :

$$BIC = ln(n)k - 2ln(\hat{L}) \tag{2.3}$$

  Pour $n \geq 8$ le terme de pénalité du BIC est supérieur à celui du AIC, favorisant

ainsi les modèles à faible dimensionnalité.

- le Bayesian Dirichlet (BD) : c'est une généralisation du **K2**. Le score Bayesian Dirichlet a été proposé par Heckerman et al. en 1995 ([17]) et est défini par :

$$BDE = \sum_{i=1}^{n} \left[ \sum_{j=1}^{q_i} \left[ log \left( \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \right) + \sum_{k=1}^{r_i} log \left( \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \right) \right] \right] \quad (2.4)$$

où $\alpha_{ijk}$ correspond à une distribution *a priori* des paramètres. En pratique ces hyperparamètres ne sont pas triviaux à définir, c'est pourquoi les auteurs ont établi une fonction de score appelée **Bayesian Dirichlet equivalent**. En supposant l'équivalence de la vraisemblance, le calcul de $\alpha_i jk$ devient simple, en introduisant le nombre d'échantillon **équivalent** $N'$ :

$$\alpha_{ijk} = N' * P(X_i = x_k, pa(X_i = x_j) | \mathcal{G}_c) \quad (2.5)$$

Le nombre $N'$ correspond en réalité au poids assigné à la distribution *a priori* comparée au nombre d'échantillons. Plus cette valeur est grande, plus l'algorithme conservera la connectivité des graphes retournés par la méthode sera importante. L'évaluation de ce score sur les réseaux benchmarks ([2]) a été réalisée avec $N' = 10$, ce qui correspond à un standard de la littérature.

Comme soulevé précédemment, l'utilisation de cette famille d'algorithmes présente plusieurs limitations majeures : la première provenant du fait qu'elle est limitée à la reconstruction de graphes dits "bayésiens", c'est à dire acycliques et complètement orientés : excluant de fait les liens bi-dirigés ($\leftrightarrow$) - qui indiquent, selon le formalisme classique de la théorie des graphes, la probable présence d'une ou de plusieurs causes communes non observées - mais également les liens non dirigés, pouvant être expliqués par l'inverse des variables latentes (causes communes), appelées variables de sélection.
Une seconde limite réside dans la croissance super exponentielle de la taille de l'espace de tous les réseaux possibles en fonction du nombre de variables requérant un temps de calcul trop important pour l'examen de l'ensemble des configurations possibles au-delà de 10 ou 15 variables. Le fait que les solutions heuristiques ne garantissent pas la découverte de la configuration optimale découle directement de cette caractéristique.

CHAPITRE 3

# MÉTHODES BASÉES SUR LA CONTRAINTE : EXEMPLE DE L'ALGORITHME PC

Cette seconde classe de méthodes a pour but de découvrir progressivement les indépendances structurelles conditionnelles entre les variables (donc les nœuds) par l'intermédiaire d'un test statistique. Ces algorithmes renvoient un Graphe Acyclique Dirigé Partiellement Complété (CPDAG), en supposant que le DAG sous-jacent respecte les hypothèses de fidélité et de suffisance causale. L'utilisation du test statistique pour déterminer l'ensemble des indépendances conditionnelles implique en général la définition d'un hyper paramètre, le seuil $\alpha$. Ces approches réalisent la reconstruction en un temps polynomial sur des graphes sous-jacents peu connectés et sont applicables à des problèmes de reconstruction de grande taille. Ce chapitre présente le principe général de ces méthodes en prenant l'exemple de PC, un algorithme largement utilisé mis au point par Peter Spirtes et Clark Glymour (le nom "PC" fait donc référence aux initiales de leurs prénoms) en 1991 ([13], [41]).

## 3.1 Exemple de l'algorithme PC

La reconstruction de modèles graphiques par PC s'effectue en trois étapes distinctes :

1. l'apprentissage du squelette du réseau, par découverte des indépendances conditionnelles (grâce à un test statistique d'indépendance, selon l'hyper paramètre $\alpha$)
2. l'orientation des **v-structures** du modèle reconstruit durant la première étape
3. la **propagation** de ces orientations en suivant les règles $R_{1:3}$, dont le but est d'éviter l'ajout de nouvelles v-structures, ainsi que de cycles dirigés.

### 3.1.1 Reconstruction du squelette

L'apprentissage du squelette du réseau commence avec un graphe non dirigé, entièrement connecté (**graphe complet**). La méthode consiste à considérer l'ensemble des paires de nœuds possibles et à déterminer si ces paires sont indépendantes lorsqu'elles sont conditionnées sur un certains de leurs voisins (c'est **l'ensemble de séparation**). On commence avec une taille d'ensemble de séparation nulle : on recherche des paires de variables **marginalement indépendantes**, c'est-à-dire n'ayant pas besoin de conditionnement pour être déclarées indépendantes (on note $X \perp\!\!\!\perp Y|\emptyset$). Lorsque de telles indépendances sont découvertes, le lien correspondant est supprimé du graphe l'ensemble de séparation $\emptyset$ est stocké.
On incrémente ensuite la taille de l'ensemble de conditionnement et chaque lien $X_i, X_j$ restant est testé en réalisant un conditionnement successif sur tous les **voisins** de $X_i$ et/ou de $X_j$. De nouveau, les liens pour lesquels une indépendance conditionnelle est découverte sont supprimés, et les ensemble de séparation correspondants sont stockés. La taille des ensembles de conditionnement est incrémenté, jusqu'à ce que tous les ensembles d'adjacence considérés soient plus petit que cette taille. Le squelette correspondant, ainsi que la liste des ensemble de séparation est retourné une fois ce critère d'arrêt atteint.

### 3.1.2 Orientation du squelette

Une fois l'ensemble des indépendances conditionnelles inférées, il est possible d'orienter le squelette obtenu, en deux étapes distinctes : l'orientation des v-structures et la propagation.

#### Orientation des v-structures

Cette étape consiste à considérer un à un les **triplets ouverts** du squelette reconstruit lors de la première étape et de déterminer s'ils sont des v-structures. Considérons alors un triplet ($\langle X_i, X_j, X_k \rangle$) ouvert, constitué des liens $X_i - X_k$ et $X_k - X_j$. Si $X_k$ n'est la cause ni de $X_i$, ni de $X_j$, il n'a pas été nécessaire de conditionner sur $X_k$ pour conclure à l'indépendance conditionnelle entre $X_i$ et $X_j$ ; ce qui veut dire qu'en supposant les hypothèses de fidélité et de suffisance causale, la seule orientation possible de ce triplet est de la forme d'une v-structure. Autrement dit, un tel triplet ouvert est une v-structure si et seulement si $X_k$ n'appartient pas à l'ensemble de séparation de la paire $(X_i, X_j)$. On oriente donc les liens correspondants à ces cas durant cette étape.

### Propagation des orientations

Cette dernière étape consiste à **propager** les orientations précédemment découvertes, en faisant deux hypothèses :

1. toutes les v-structures ont été découvertes
2. le graphe sous-jacent est un DAG et ne contient par conséquent aucun cycle dirigé

En suivant ces deux hypothèses, ainsi que les règles d'orientation correspondantes ($R_{1:3}$), on peut propager les orientations des v-structures précédemment définies.

---

**Algorithme 1 :** Algorithme PC

---

**Entrée :** Données d'observation de l'ensemble de variables **O**, classement des variables de **O**, seuil de significativité $\alpha$
**Sortie :** CPDAG $\mathcal{C}$

**0. Initialisation**
Débuter avec un graph complet non dirigé $\mathcal{G}$
Initialiser $l = 0$
**1. Itération**
**tant que** $\exists X \in \mathcal{G}$ *tel que* $|adj_{\mathcal{G}} \geq l$ **faire**
  **tant que** $\exists X_i X_j \subset adj_{\mathcal{G}}$ *et* $\exists \{U_i\} \subseteq adj_{\mathcal{G}}(X_i) \backslash \{X_j\}$ *non considéré, avec* $|\{U_i\}| = l$ **faire**
    **si** $X_i \perp\!\!\!\perp X_j | \{U_i\}$ **alors**
      **le lien** $X_i X_j$ n'est pas essentiel et est **supprimé**
      **set de séparation** de $X_i X_j$ : $\mathrm{Sep}_{X_i X_j} = \{U_i\}$
  **fin**
  $l = l + 1$
**fin**

**2. Orientation**
**pour tous** *les triplets ouverts* **faire**
  $R_0 : \{X_i - X_k - X_j \;\&\; \nexists\, X_i - X_j \;\&\; X_k \notin Sep_{X_i X_j}\} \Rightarrow \{X_i \rightarrow X_k \leftarrow X_j\}$
**fin**

**3. Propagation**
**tant que** $\exists$ *une orientation pouvant être propagée* **faire**
  $R_1 : \{X_i \rightarrow X_k - X_j \;\&\; \nexists\, X_i - X_j \;\Rightarrow \{X_i \rightarrow X_k \rightarrow X_j\}$
  $R_2 : \{X_i \rightarrow X_k \rightarrow X_j \;\&\; X_i - X_j\} \Rightarrow \{X_i \rightarrow X_j\}$
  $R_3 : \{X_i - X_k \rightarrow X_j \;\&\; X_i - X_l \rightarrow X_j \;\&\; \nexists\, X_k - X_l\} \Rightarrow \{X_i \rightarrow X_j\}$
**fin**

---

### 3.1.3 Tests statistiques pour les indépendances conditionnelles

La découverte des indépendances conditionnelles est généralement liée à des tests statistiques de type $\chi^2$ ou $G^2$. Ils permettent, en utilisant le seuil de significativité $\alpha$, de tester les deux hypothèses :

- $H_0$ : les deux variables sont indépendantes lorsqu'elles sont conditionnées sur l'ensemble de séparation considéré ($p - value > \alpha$)
- $H_1$ : les deux variables sont dépendantes lorsqu'elles sont conditionnées sur l'ensemble de séparation considéré ($p - value < \alpha$)

#### Test d'indépendance conditionnelle du $\chi^2$

Pour accepter ou rejeter l'hypothèse nulle définies ci-dessus, on peut utiliser le test du $\chi^2$ sur des variables discrètes. Si on note $X_i$, $X_j$ les deux variables considérées et $\{X_k\}$ l'ensemble de variables sur lesquelles on les conditionne. Le but est de comparer les deux modèles suivants :

- le modèle ou les deux variables sont indépendantes : $p_i = P(X_i|X_k)P(X_j|X_k)$
- le modèle observé : $p_0 = P(X_i, X_j|X_k)$

Le modèle $p_0$ est représenté par $N_{ijk}$, le nombre d'échantillons (avec $X_i = x_i,\ X_j = x_j$ et $X_k = x_k$ où $x_k$ correspond à une combinaison $k$ de l'ensemble $X_k$). Le modèle $p_i$ est quant à lui représenté par $\frac{N_{ik} * N_{jk}}{N_k}$ où $N_{ik}$, $N_{jk}$ et $N_k$ sont le nombre d'échantillons des données respectant, respectivement $\{X_i = x_i$ et $X_k = x_k\}$, $\{X_j = x_j$ et $X_k = x_k\}$ et $\{X_k = x_k\}$. La statistique du test est définie de la manière suivante :

$$\chi^2_{statistics} = \sum_{i,j,l}^{r_i} \sum_{j=1}^{r_j} \sum_{k=1}^{q_k} \frac{(N_{ijk} - \frac{N_{ik} * N_{jk}}{N_k})^2}{\frac{N_{ik} * N_{jk}}{N_k}} \tag{3.1}$$

où $r_i$ et $r_j$ correspondent respectivement au nombre de niveau de $X_i$ et $X_j$ et où $q_k$ correspond au nombre de combinaisons de niveaux pour l'ensemble $X_k$. L'indépendance entre $X_i$ et $X_j$ conditionnées sur $X_k$ est alors rejetée pour un seuil de significativité $\alpha$ si $P(\chi^2_{statistics} < \chi^2(ddl))$ avec $ddl = (r_i - 1)(r_j - 1)q_k$.

**Test d'indépendance conditionnelle du $G^2$**

Il est également possible d'utiliser le ratio de vraisemblance $G^2$, comme l'on proposé Spirtes *et al.*. Celui-ci prend la forme d'un $\chi^2$ avec $ddl = (r_i - 1)(r_j - 1)q_k$ :

$$G^2 = 2\sum_{i=1}^{r_i}\sum_{j=1}^{r_j}\sum_{k=1}^{q_k} N_{ijk} ln\left(\frac{N_{ijk}}{\frac{N_{ik} * N_{jk}}{N_k}}\right) \tag{3.2}$$

De la même manière que pour le $\chi^2$, ce ratio de vraisemblance $G^2$ peut être utilisé pour rejeter ou accepter l'hypothèse nulle avec un risque $\alpha$.

**Effet du seuil $\alpha$**

La définition de l'hyper paramètre $\alpha$ a un impact important sur la reconstruction du réseau par l'algorithme PC. En effet, un seuil élevé a tendance à moins conclure à une indépendance entre les variables (les p-values ont plus de chance d'être $< 1$), ce qui favorise un réseau plus connecté, avec un risque d'inférer des liens **faux positifs** de plus en plus important. A l'inverse, un $\alpha$ petit rejette plus rarement l'hypothèse $H_0$ ; ce qui favorise la suppression des liens, avec un risque d'inférence de liens **faux négatifs** de plus en plus important.

## 3.2 Améliorations de l'algorithme PC

### 3.2.1 PC-Stable

Il a été démontré ([5]) que les méthodes basées sur la contrainte sont peu robustes au bruit dû à l'échantillonnage en raison de l'utilisation d'un test statistique sur un jeu de données fini. En effet, les erreurs commises en début d'exécution de l'algorithme ont tendance à entraîner une accumulation d'erreurs compensatoires durant la suite du déroulement de l'algorithme. Ces cascades d'erreurs rendent ces méthodes vulnérables aux variations du seuil $\alpha$. De plus, elles sont également sensibles à l'ordre dans lequel les liens sont testés, puisqu'il influence les ensembles d'adjacences testés comme ensembles de séparation. En effet, lorsqu'un lien entre deux variables $X_i$ et $X_j$ est supprimé, le conditionnement de l'une sur l'autre devient interdit par l'algorithme, puisque $X_i \notin adj_{\mathcal{G}}(X_j)$ et $X_j \notin adj_{\mathcal{G}}(X_i)$.
Ceci rend les résultats de PC très dépendants de l'ordre dans lequel sont considérés les

liens, puisque chaque ordre entraîne potentiellement des erreurs différentes. Cette problématique est adressée par l'algorithme **PC stable**, une amélioration de l'algorithme d'origine proposée par Colombo et Maathuis en 2014([5]). Les auteurs proposent en effet de **retarder** la suppression des liens, en conservant en mémoire les indépendances conditionnelles découvertes pour une taille de séparation $l$ donnée (voir pseudocode 1) : ce n'est que lorsque tous les ensembles de conditionnement correspondant à cette taille $l$ ont été testés que les liens pour lesquels une indépendance conditionnelle à été découverte sont supprimés. Grâce à cette modification relativement simple, les résultats de l'inférence du squelette sont indépendants de l'ordre de test des liens.

### 3.2.2 PC Conservative

Il a également été démontré que le classement des variables a un impact sur la détermination des v-structures, au moment de l'étape 2 de l'algorithme 1. En effet, en raison de ce classement, des squelettes identiques basés sur des ensembles de séparation différents peuvent être inférés (exemple 2, [6]), ce qui peut éventuellement mener à l'inférence de v-structures erronées.

Une modification de l'algorithme PC, appelée **PC conservative**, a été proposée par Ramsey et al en 2006 ([33]) pour éviter cette problématique. Plus tard, en 2012, Colombo et Maathuis ([6]) ont proposé une légère modification de cette règle dans le but de la rendre plus flexible, afin de permettre l'identification d'un plus grand nombre de v-structures. Ces deux modifications sont basées sur l'identification de triplet non-ambigus et ne diffèrent que par la définition de non-ambiguïté. Tel que défini par Ramsey et al, un triplet ouvert $\langle X_i, X_k, X_j \rangle$ de la forme $X_i - X_k - X_j$ est non-ambigu si au moins un set de séparation $Y$ est trouvé entre $X_i$ et $X_j$ (c'est-à-dire que $X_i \perp\!\!\!\perp X_j | Y$), et si $X_k$ appartient à tous (ou à aucun) ensemble de séparation $Y$. Le triplet non ambigu est donc orienté comme une v-structure si et seulement si $X_k$ ne se trouve dans **aucun** des ensembles de séparation $Y$ découvert.

Colombo et Maathuis proposent quant à eux une version moins contraignante de l'ambiguïté. En effet, ils définissent un triplet ouvert $\langle X_i, X_k, X_j \rangle$ de la forme $X_i - X_k - X_j$ comme étant non-ambigu si au moins un ensemble de séparation $Y$ est trouvé et si la proportion de ces ensembles contenant $X_k$ est différente de $50\%$. Un triplet non ambigu tel que défini de cette manière est orienté comme une v-structure si et seulement si $X_k$ est absent de plus de $50\%$ des ensembles de séparation trouvés entre $X_i$ et $X_j$. Cette règle est appelée la **règle de la majorité**. Les algorithmes résultant de l'application de ces deux

règles sont respectivement appelés **CPC-stable** et **MPC-stable**.

### 3.2.3 Signalement des conflits d'orientation

La dernière amélioration de l'algorithme PC concerne les conflits d'orientation entre deux v-structures. En effet, il est théoriquement possible que la méthode trouve deux v-structures contradictoires, du type $X_i \rightarrow X_k \leftarrow X_j$ et $X_k \rightarrow X_j \leftarrow X_m$. Ce conflit peut être signalé et la variabilité de la v-structure inféré en fonction de l'ordre des triplets peut être évité grâce à un lien bi-dirigé entre $X_k$ et $X_j$. Il est cependant important de noter que ces liens bi-dirigés ne peuvent être considérés comme étant causaux ([5]) ; ils ne sont donc que les indicateurs du conflit d'orientation et ne sont pas à interpréter comme reflétant une variable latente (voir section 5.4.2).

CHAPITRE 4

# 3OFF2 : BASÉ SUR LES CONTRAINTES ET SUR LA THÉORIE DE L'INFORMATION

## 4.1 Contexte et principe

Ce paragraphe a pour but d'introduire le cadre de théorie de l'information sur lequel est basé `3off2`, première version publiée de l'algorithme de reconstruction de modèles graphiques mise au point par l'équipe Isambert.
Tout comme PC, `3off2` part d'un réseaux complet (toutes les variables sont directement connectées entre elles) et teste chaque lien afin de vérifier s'il est essentiel pour expliquer les données observées. `3off2` procède ensuite à une étape d'orientation/propagation permettant d'inférer la causalité inhérentes aux données, suivant ainsi le fonctionnement classique des méthodes basées sur la contrainte.
La nouveauté de cette méthode réside dans l'utilisation d'une fonction de score pour la découverte des indépendances conditionnelles. Rappelons en effet que pour un lien $X_i, X_j$, le but principal de la reconstruction de réseaux est de différencier une interaction directe entre les deux variables $X_i$ et $X_j$ d'une interaction indirecte, ayant une ou plusieurs variables intermédiaire : il faut donc découvrir les variables susceptibles de jouer ce rôle dans le système observé (indépendances conditionnelles). Comme introduit précédemment, les méthodes classiques basées sur la contrainte recherchent arbitrairement ces contributeurs parmi les voisins de $X_i$ et/ou de $X_j$. Pour ce faire, des tests sont réalisés sur l'ensemble des combinaisons possibles de ces voisins, soit jusqu'à l'obtention d'un résultat permettant la suppression du lien, soit jusqu'à ce que toutes les combinaisons aient été testées et que de fait le lien soit avéré. La combinatoire exigée par cette technique pose problème, car elle augmente très rapidement en fonction du nombre de voisins.
`3off2`, recherche quant à lui uniquement les **meilleurs contributeurs** en leur attribuant

un score permettant de dégager un ensemble de séparation optimum, sans avoir besoin de tester les autres. Le score utilisé correspond (noté $R$, détaillé dans Affeldt et al., *BMC Bioinformatics*, 2016 [2]) en pratique au minimum de deux composantes :

- probabilité que la paire considérée $X_i, Xj$ et le contributeur potentiel (noté $X_k$) forment une non-v-structure : $P_{nv}(X_i, X_k, X_j)$
- probabilité que le lien testé $(X_i, X_j)$ soit la **base** de cette non-v-structure : $P_b(X_i, X_j)$.

Les contributions des meilleurs contributeurs ainsi sélectionnés sont successivement soustraites à l'information mutuelle $I'(X_i; X_j)$ ($I'(X_i; X_j) = I(X_i; X_j) - k(X_i, X_j)$, avec $k(X_i, X_j)$ terme de complexité utilisé pour prendre en compte le caractère fini du jeu de données voir [2] ), isolant ainsi l'information résiduelle $I'(X_i; X_j|\{A_i\}_n)$ :

$$I'(X_i; X_j|\{A_i\}) = I'(X_i; X_j) - I'(X_i; X_j; \{A_1\}) - I'(X_i; X_j; \{A_2\}|\{A_1\}) - ... \\ -I'(X_i; X_j; \}A_n\}|\{A_i\}_{n-1}) \tag{4.1}$$

Ces soustractions sont effectuées jusqu'à ce que $I(X_i; X_j|\{A_i\}_n) \leq 0$, auquel cas le lien est supprimé, ou qu'aucun contributeur significatif ne soit dégagé, auquel cas le lien est conservé dans le modèle final. L'étape d'orientation de `3off2` peut se décomposer de la même manière que pour les méthodes basées sur la contraintes classiques : les v-structures sont tout d'abord orientées, puis leurs orientations sont propagées lorsque c'est possible. Pour repérer les v-structures, la méthode se base sur le signe de l'information mutuelle à 3 points du triplet considéré ; ce que nous allons expliquer de façon intuitive dans le paragraphe suivant (Pour une preuve formelle, voir l'article [2]).

## 4.2 Orientation des v-structures dans `3off2`

Pour rappel, la quantité utilisée pour déterminer une indépendance conditionnelle en terme d'information mutuelle est l'information mutuelle conditionnée sur les **contributeurs**. Lorsque cette quantité est supérieure à 0, une part de l'information mutuelle entre nos deux variables n'est pas expliquée par ses contributeurs, ce qui signifie qu'elle ne peut provenir que de l'interaction **directe** entre elles. Afin d'obtenir cette information résiduelle, on soustrait l'information à 3 point qui représente la part d'information médiée par un contributeur défini. Sachant que l'information à 3 point peut être **positive ou négative**, si on cherche à soustraire une information négative à notre information mutuelle, l'information résiduelle obtenue **augmente**. De la même manière que le conditionnement

sur la pointe d'une v-structure rend dépendante deux variables marginalement indépendantes, la soustraction d'une information mutuelle à trois points négative rend l'information mutuelle conditionnée plus importante ; c'est donc le signe que l'on est en présence d'une v-structure.
En résumé, la méthode `3off2` applique le schéma décisionnel suivant pour chaque triplet ouvert de la forme $X_i - X_k - X_j$ : si $I'(X_i; X_k; X_j|\{A_i\}) < 0$ et $I'(X_i; X_j|\{A_i\}) < 0$, alors $X_i \rightarrow X_k \leftarrow X_j$.

**Algorithme 1 :** Reconstruction d'un modle graphique par `3off2`

**Entre:** Donnes d'observation de taille finie $N$, critre de complexit NML
**Sortie:** Rseau (partiellement) orient $\mathcal{G}$

**0. Initialisation**
Dbuter avec un graph complet non dirig $\mathcal{G}$
**pour tous** *les liens* $X_iX_j$ **faire**
  **si** $I'(X_i;X_j)<0$ **alors**
    **le lien** $X_iX_j$ n'est pas essentiel et est **supprim**
    **set de sparation** de $X_iX_j$: $\mathrm{Sep}_{X_iX_j}=\emptyset$
  **sinon**
    trouver le **meilleur contributeur** $X_k$ parmi les voisins de $X_i$ ou de $X_j$ et **calculer son rang**, $R(X_iX_j;Z|\emptyset)$
  **fin**
**fin**

**1. Itration**
**tant que** $\exists$ *un lien* $X_iX_j$ *avec* $R(X_iX_j;X_k|\{A_i\})>1/2$ **faire**
  **pour** *lien* $X_iX_j$ *avec le meilleur rang* $R(X_iX_j;X_k|\{A_i\})$ **faire**
    **mettre jour l'ensemble des contributeurs** $\{A_i\}\leftarrow\{A_i\}+Z$
    **si** $I'(X_i;X_j|\{A_i\})<0$ **alors**
      **le lien** $X_iX_j$ n'est pas essentiel et est **supprim**
      **set de sparation** de $X_iX_j$: $\mathrm{Sep}_{X_iX_j}=\{A_i\}$
    **sinon**
      trouver le **meilleur contributeur suivant** $X_k$, voisin de $X_i$ ou de $X_j$ et **calculer son rang**, $R(X_iX_j;X_k|\{A_i\})$
    **fin**
    **mettre jour la liste des rangs** $R(X_iX_j;X_k|\{A_i\})$
  **fin**
**fin**

**2. Orientation/Propagation**
Classer la liste des triplets ouverts, $\mathcal{L}_c=\{\langle X_i,X_j,X_k\rangle_{\nexists\ X_i-X_j}\}$ en ordre dcroissant selon $|\ I'(X_i;X_j;X_k\mid u_i)\ |$
**tant que** $\exists$ *une orientation additionnelle possible* **faire**
  Considrer $\langle X_i;X_j;X_k\rangle_{\nexists\ X_i-X_j}\in\mathcal{L}_c$ avec la plus grande valeur de $|\ I'(X_i;X_j;X_k\mid u_i)\ |$ sur lequel les rgles $R_0$ ou $R_1$ peuvent łtre appliques
  **si** $I'(X_i;X_j;X_k\mid u_i)<0$ **alors**
    si $\langle X_i,X_j,X_k\rangle_{\nexists\ X_i-X_j}$ ne prsente pas d'orientation divergente, appliquer:
    $R_0:\{X_i-*X_k*-X_j\ \&\ \nexists\ X_i-X_j\ \&\ X_k\notin Sep_{X_iX_j}\}\Rightarrow\{X_i\rightarrow X_k\leftarrow X_j\}$
  **sinon**
    si $\langle X_i,X_k,X_j\rangle_{\nexists\ X_i-X_j}$ prsente une orientation convergente, appliquer:
    $R_1:\{X_i\rightarrow X_k-X_j\ \&\ \nexists\ X_i-X_j\}\Rightarrow\{X_k\rightarrow X_j\}$
  **fin**
**fin**

## 4.3 Publication de `3off2` dans *BMC Bioinformatics*

Affeldt *et al. BMC Bioinformatics* 2016, **17**(Suppl 2):12
DOI 10.1186/s12859-015-0856-x

BMC Bioinformatics

PROCEEDINGS Open Access

# 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics

Séverine Affeldt[1,2], Louis Verny[1,2] and Hervé Isambert[1,2*]

*From* Bringing Maths to Life (BMTL)
Naples, Italy. 27-29 October 2014

## Abstract

**Background:** The reconstruction of reliable graphical models from observational data is important in bioinformatics and other computational fields applying network reconstruction methods to large, yet finite datasets. The main network reconstruction approaches are either based on Bayesian scores, which enable the ranking of alternative Bayesian networks, or rely on the identification of structural independencies, which correspond to missing edges in the underlying network. Bayesian inference methods typically require heuristic search strategies, such as hill-climbing algorithms, to sample the super-exponential space of possible networks. By contrast, constraint-based methods, such as the PC and IC algorithms, are expected to run in polynomial time on sparse underlying graphs, provided that a correct list of conditional independencies is available. Yet, in practice, conditional independencies need to be ascertained from the available observational data, based on adjustable statistical significance levels, and are not robust to sampling noise from finite datasets.

**Results:** We propose a more robust approach to reconstruct graphical models from finite datasets. It combines constraint-based and Bayesian approaches to infer structural independencies based on the ranking of their most likely contributing nodes. In a nutshell, this local optimization scheme and corresponding **3off2** algorithm iteratively "take off" the most likely conditional 3-point information from the 2-point (mutual) information between each pair of nodes. Conditional independencies are thus derived by progressively collecting the most significant indirect contributions to all pairwise mutual information. The resulting network skeleton is then partially directed by orienting and propagating edge directions, based on the sign and magnitude of the conditional 3-point information of unshielded triples. The approach is shown to outperform both constraint-based and Bayesian inference methods on a range of benchmark networks. The **3off2** approach is then applied to the reconstruction of the hematopoiesis regulation network based on recent single cell expression data and is found to retrieve more experimentally ascertained regulations between transcription factors than with other available methods.

**Conclusions:** The novel information-theoretic approach and corresponding **3off2** algorithm combine constraint-based and Bayesian inference methods to reliably reconstruct graphical models, despite inherent sampling noise in finite datasets. In particular, experimentally verified interactions as well as novel predicted regulations are established on the hematopoiesis regulatory networks based on single cell expression data.

**Keywords:** Network reconstruction, Hybrid inference method, Information theory, Hematopoiesis

*Correspondence: herve.isambert@curie.fr
[1] Institut Curie, PSL Research University, CNRS, UMR168, 26 rue d'Ulm, 75005 Paris, France
[2] Sorbonne Universités, UPMC Univ Paris 06, 4, Place Jussieu, 75005 Paris, France

## Background

Two types of reconstruction method for directed networks have been developed and applied to a variety of experimental datasets. These methods are either based on Bayesian scores [1, 2] or rely on the identification of structural independencies, which correspond to missing edges in the underlying network [3, 4].

Bayesian inference approaches have the advantage of allowing for quantitative comparisons between alternative networks through their Bayesian scores but they are limited to rather small causal graphs due to the super-exponential space of possible directed graphs to sample [1, 5, 6]. Hence, Bayesian inference methods typically require either suitable prior restrictions on the structures [7, 8] or heuristic search strategies such as hill-climbing algorithms [9–11].

By contrast, structure learning algorithms based on the identification of structural constraints typically run in polynomial time on sparse underlying graphs. These so-called constraint-based approaches, such as the PC [12] and IC [13] algorithms, do not score and compare alternative networks. Instead they aim at ascertaining conditional independencies between variables to directly infer the Markov equivalent class of all causal graphs compatible with the available observational data. Yet, these methods are not robust to sampling noise in finite datasets as early errors in removing edges from the complete graph typically trigger the accumulation of compensatory errors later on in the pruning process. This cascading effect makes the constraint-based approaches sensitive to the adjustable significance level $\alpha$, required for the conditional independence tests. In addition, traditional constraint-based methods are not robust to the order in which the conditional independence tests are processed, which prompted recent algorithmic improvements intending to achieve order-independence [14].

In this paper, we report a novel network reconstruction method, which exploits the best of these two types of structure learning approaches. It combines constraint-based and Bayesian frameworks to reliably reconstruct graphical models despite inherent sampling noise in finite observational datasets. To this end, we have developed a robust information-theoretic method to confidently ascertain structural independencies in causal graphs based on the ranking of their most likely contributing nodes. Conditional independencies are derived using an iterative search approach that identifies the most significant indirect contributions to all pairwise mutual information between variables. This local optimization algorithm, outlined below, amounts to iteratively subtracting the most likely conditional 3-point information from 2-point information between each pair of nodes. The resulting network skeleton is then partially directed by orienting and propagating edge directions, based on the sign and magnitude of the conditional 3-point information of unshielded triples. Identifying structural independencies within such a maximum likelihood framework circumvents the need for adjustable significance levels and is found to be more robust to sampling noise from finite observational data, even when compared to constraint-based methods intending to resolve the order-dependence on the variables [14].

### Constraint-based methods

Constraint-based approaches, such as the PC [12] and IC [13] algorithms, infer causal graphs from observational data, by searching for conditional independencies among variables. Under the Markov and Faithfulness assumptions, these algorithms return a Complete Partially Directed Acyclic Graph (CPDAG) that represents the Markov equivalent class of the underlying causal structure [3, 4]. They proceed in three steps detailed in Algorithm 1:

**Algorithm 1:** Constraint-based network reconstruction

**In:** observational data of variables $\mathbf{V}$; an ordering $order(\mathbf{V})$ on the variables; a significance level $\alpha$
**Out:** CPDAG $\mathcal{C}$
**0. Initiation**
Start with a complete undirected graph $\mathcal{G}$
Let $\ell = 0$
**1. Iteration**
**repeat**
  **while** $\exists$ *xy link with* $|adj(\mathcal{G},x)\setminus\{y\}| \geqslant \ell$ **do**
    **while** $xy \subset adj(\mathcal{G})$ *and* $\exists\{u_i\} \subseteq adj(\mathcal{G},x)\setminus\{y\}$, *not yet considered with* $|\{u_i\}| = \ell$ **do**
      **if** $Indep(x;y|\{u_i\})$ *at significance level* $\alpha$ **then**
        ***xy* link is** non-essential and **removed**
        **separation set** of $xy$: $\mathrm{Sep}_{xy} = \{u_i\}$
      **end**
    **end**
  **end**
  **Set** $\ell = \ell + 1$
**until** $\forall x \in \mathcal{G}$, $|adj(\mathcal{G},x)| \leqslant \ell$;
**2. Orientation**
**forall the** *unshielded triples* **do**
  $R_0$: $\{x - z - y \;\&\; x \not\to y \;\&\; z \notin \mathrm{Sep}_{xy}\} \Rightarrow \{x \to z \leftarrow y\}$
**end**
**3. Propagation**
**repeat**
  $R_1$: $\{x \to z - y \;\&\; x \not\to y\} \Rightarrow \{z \to y\}$
  $R_2$: $\{x \to y \to z \;\&\; x - z\} \Rightarrow \{x \to z\}$
  $R_3$: $\{x - y \to z \;\&\; x - t \to z \;\&\; y \not\to t\} \Rightarrow \{x \to z\}$
**until** *no further orientation can be propagated*;

- 1) inferring unnecessary edges and associated separation sets to obtain an undirected skeleton.
- 2) orienting unshielded triples as v-structures if their middle node is not in the separation set ($R_0$).
- 3) propagating as many orientations as possible following propagation rules ($R_{1–3}$), which prevents the orientation of additional v-structures ($R_3$) and directed cycles ($R_{2–3}$) [15].

However, as previously stated, the sensitivity of the constraint-based methods to the adjustable significance level $\alpha$ used for the conditional independence tests and to the order in which the variables are processed (step 1) favors the accumulation of errors when the search procedure relies on finite observational data.

In this paper, we aim at improving constraint-based methods, Algorithm 1, by uncovering the most reliable conditional independencies supported by the (finite) available data, based on a quantitative information theoretic framework.

#### Maximum likelihood methods

The maximum likelihood $\mathcal{L}_{\mathcal{G}}$ is related to the cross entropy $H(\mathcal{G},\mathcal{D}) = -\sum_{\{x_i\}} p(\{x_i\}) \log(q(\{x_i\}))$ between the "true" probability distribution $p(\{x_i\})$ from the data $\mathcal{D}$ and the approximate probability distribution $q(\{x_i\}) = \prod_i p(x_i|\{Pa_{x_i}\})$ generated by the Bayesian network $\mathcal{G}$ with specific parent nodes $\{Pa_{x_i}\}$ for each node $x_i$, leading to [16],

$$\mathcal{L}_{\mathcal{G}} = e^{-NH(\mathcal{G},\mathcal{D})} = e^{-N\sum_i H(x_i|\{Pa_{x_i}\})} \tag{1}$$

where $\sum_i H(x_i|\{Pa_{x_i}\})$ is the (conditional) entropy of the underlying causal graph. This enables to score and compare alternative models through their maximum likelihood ratio as,

$$\frac{\mathcal{L}_{\mathcal{G}'}}{\mathcal{L}_{\mathcal{G}}} = e^{-N\sum_i \left(H\left(x_i|\left\{Pa'_{x_i}\right\}\right) - H(x_i|\{Pa_{x_i}\})\right)} \tag{2}$$

Note, in particular, that the significance level of the Maximum likelihood approach is set by the number $N$ of independent observational data points, as detailed in the Methods Section below.

## Methods

### Information theoretic framework

#### *Inferring isolated v-structures vs non-v-structures from 3-point and 2-point information*

Applying the previous likelihood definition, Eq. 1, to isolated v-structures (Fig. 1a) and Markov equivalent non-v-structures (Fig. 1b–d), one obtains,

$$\begin{aligned}\mathcal{L}_{\mathsf{v}}(xy) &= e^{-N\left[H(z|x,y)+H(x)+H(y)\right]} \\ &= e^{-N\left[H(x,y,z)+I(x;y)\right]}\end{aligned} \tag{3}$$

where $I(x;y) = H(x)+H(y)-H(x,y)$ is the 2-point mutual information between $x$ and $y$, and,

$$\begin{aligned}\mathcal{L}_{\mathsf{nv}}(xy) &= e^{-N\left[H(x|z)+H(y|z)+H(z)\right]} \\ &= e^{-N\left[H(x,y,z)+I(x;y|z)\right]}\end{aligned} \tag{4}$$

where $I(x;y|z) = H(x|z)+H(y|z)-H(x,y|z)$ is the conditional mutual information between $x$ and $y$ given $z$. Hence, one obtains the likelihood ratio,

$$\frac{\mathcal{L}_{\mathsf{v}}(xy)}{\mathcal{L}_{\mathsf{nv}}(xy)} = e^{-N\left[I(x;y)-I(x;y|z)\right]} = e^{-NI(x;y;z)} \tag{5}$$

where we introduced the 3-point information function, $I(x;y;z) = I(x;y) - I(x;y|z)$, which is in fact invariant upon permutations between $x, y$ and $z$, as seen in terms of entropy functions,

$$\begin{aligned}I(x;y;z) = {} & H(x)+H(y)+H(z)-H(x,y) \\ & -H(x,z)-H(y,z)+H(x,y,z)\end{aligned} \tag{6}$$

As long recognized in the field [17, 18], 3-point information, $I(x;y;z)$, can be positive or negative (if $I(x;y) < I(x;y|z)$), unlike 2-point mutual information, which are always positive, $I(x;y) \geqslant 0$.

More precisely, Eq. 5 demonstrates that the sign and magnitude of 3-point information provide a quantitative estimate of the relative likelihoods of isolated v-structures *versus* non-v-structures, which are in fact independent of their actual non-connected bases $xy$, $xz$ or $yz$,

$$\frac{\mathcal{L}_{\mathsf{v}}(xy)}{\mathcal{L}_{\mathsf{nv}}(xy)} = \frac{\mathcal{L}_{\mathsf{v}}(xz)}{\mathcal{L}_{\mathsf{nv}}(xz)} = \frac{\mathcal{L}_{\mathsf{v}}(yz)}{\mathcal{L}_{\mathsf{nv}}(yz)} = e^{-NI(x;y;z)} \tag{7}$$

Hence, a significantly negative 3-point information, $I(x;y;z) < 0$, implies that a v-structure is more likely than a non-v-structure given the observed correlation data. Conversely, a significantly positive 3-point information, $I(x;y;z) > 0$, implies that a non-v-structure model is more likely than a v-structure model.

Yet, as noted above, 3-point information, $I(x;y;z)$, being symmetric by construction, it cannot indicate how to orient v-structures or non-v-structures over the $xyz$ triple. To this end, it is however straightforward to show that the most likely base ($xy$, $xz$ or $yz$) of the local v-structure or non-v-structure corresponds to the pair with lowest

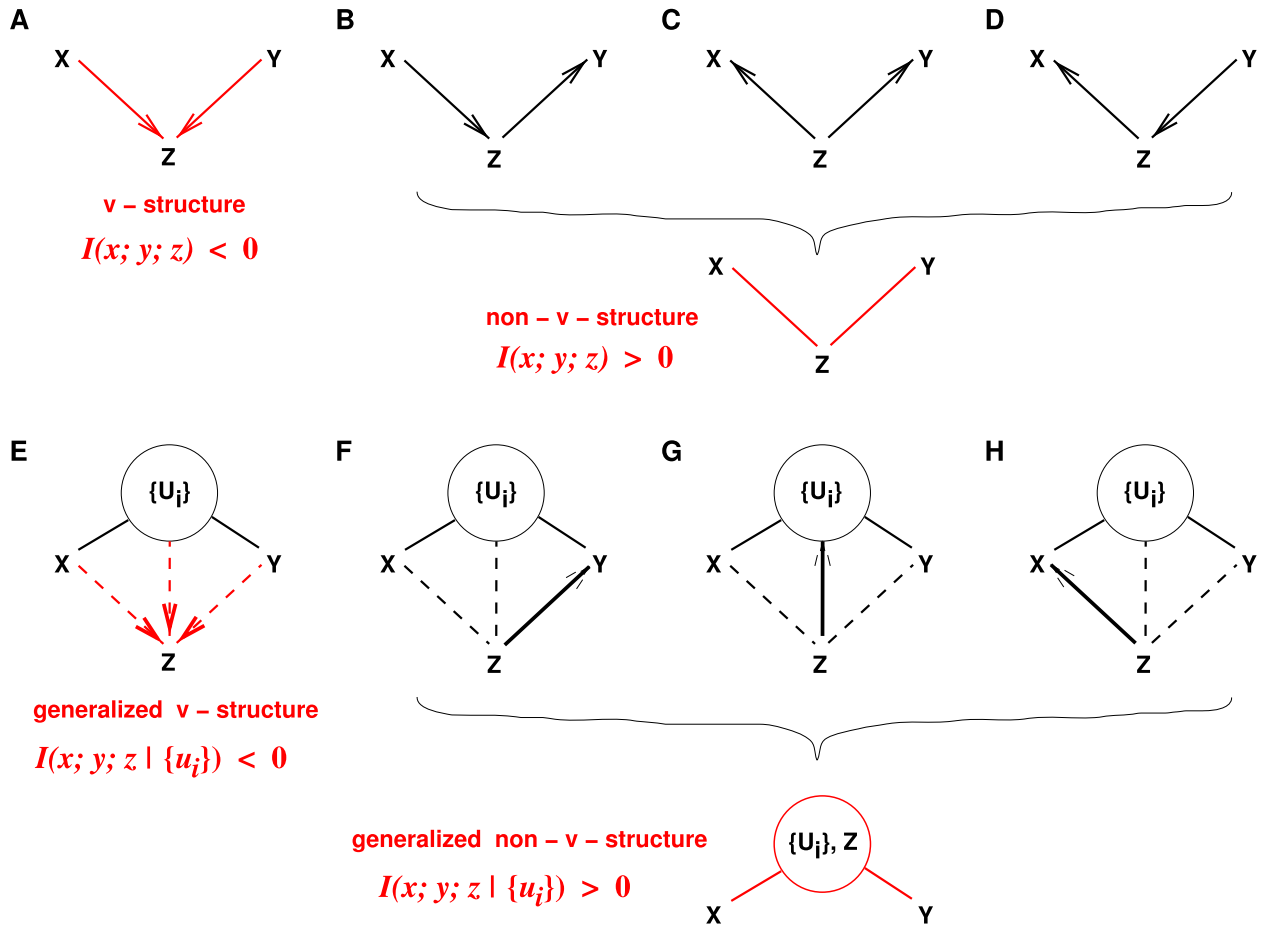



**Fig. 1** Inference of v-structures *versus* non-v-structures by 3-point information from observational data. **a** Isolated v-structures are predicted for $I(x;y;z) < 0$, and (**b**–**d**) isolated non-v-structures for $I(x;y;z) > 0$. **e** Generalized v-structures are predicted for $I(x;y;z|\{u_i\}) < 0$ and (**f**–**h**) generalized non-v-structures for $I(x;y;z|\{u_i\}) > 0$. In addition, as $I(x;y;z|\{u_i\})$ are invariant upon *xyz* permutations, the global orientation of v-structures and non-v-structures also requires to find the most likely base of the *xyz* triple. Choosing the base *xy* with the lowest conditional mutual information, *i.e.*, $I(x;y|\{u_i\}) = \min_{xyz}(I(s;t|\{u_i\}))$, is found to be consistent with the Data Processing Inequality expected for (generalized) non-v-structures in the limit of infinite dataset, see main text. In practice, given a finite dataset, the inference of (generalized) v-structures *versus* non-v-structures can be obtained by replacing 3-point and 2-point information terms $I(x;y|\{u_i\})$ and $I(x;y;z|\{u_i\})$ by shifted equivalents, $I'(x;y|\{u_i\})$ and $I'(x;y;z|\{u_i\})$, including finite size corrections, see text (Eqs. 23 & 24)

mutual information, *e.g.*, $I(x;y) = \min_{xyz}\big(I(s;t)\big)$, as shown by the likelihood ratios,

$$\frac{\mathcal{L}_{\mathrm{v}}(xy)}{\mathcal{L}_{\mathrm{v}}(st)} = \frac{\mathcal{L}_{\mathrm{nv}}(xy)}{\mathcal{L}_{\mathrm{nv}}(st)} = \frac{e^{-NI(x;y)}}{e^{-NI(s;t)}} \tag{8}$$

Note, in particular, that choosing the base with the lowest mutual information is consistent with the Data Processing Inequality expected for non-v-structures, Fig. 1b–d.

Hence, combining 3-point and 2-point information allows to determine the likelihood and the base of isolated v-structures *versus* non-v-structures. But how to extend such simple results to identify local v-structures and non-v-structures embedded within an entire graph $\mathcal{G}$?

### *Inferring embedded v-structures vs non-v-structures from conditional 3-point and 2-point information*

To go from isolated to embedded v-structures and non-v-structures within a DAG $\mathcal{G}$, we will consider the Markov equivalent CPDAG of $\mathcal{G}$ and introduce generalized v-structures and non-v-structures, Fig. 1e–h. We will demonstrate that their relative likelihood, given the available observational data, can be estimated from the sign and magnitude of a conditional 3-point information, $I(x;y;z|\{u_i\})$, Eq. 11. This will extend our initial result valid for isolated v-structures and non-v-structures, Eq. 7.

Let's consider a pair of non-neighbor nodes $x, y$ with a set of upstream nodes $\{u_i\}_n$, where each node $u_i$ has at least one direct connection to $x$ ($u_i \to x$) or $y$ ($u_i \to y$) or to another upstream node $u_j \in \{u_i\}_n$ ($u_i \to u_j$) or only undirected links to these nodes ($u_i - x$, $u_i - y$ or $u_i - u_j$). Thus, given $x, y$ and a set of upstream nodes $\{u_i\}_n$, any additional node $z$ can either be:

- *i)* at the apex of a generalized v-structure, if *all* existing connections between $x$, $y$, $\{u_i\}_n$ and $z$ are directed and point towards $z$, Fig. 1e, or else,
- *ii)* $z$ has *at least one* undirected link with $x$, $y$ or one of the upstream nodes $u_i$ ($z - x$, $z - y$ or $z - u_i$) *or at least one* directed link pointing towards these nodes ($z \to x$, $z \to y$ or $z \to u_i$), Fig. 1f–h. In such a case, $z$ might contribute to the mutual information $I(x; y)$ and should be included in the set of upstream nodes $\{u_i\}_n$, thereby defining a generalized non-v-structure, Figs. 1f–h.

Then, similarly to the case of an isolated v-structure (Eq. 3), the maximum likelihood $\mathcal{L}_{\mathsf{v}}(xy)$ of a generalized v-structure pointing towards $z$ from a base $xy$ with upstream nodes $\{u_i\}_n$ can be expressed as,

$$\begin{aligned}\mathcal{L}_{\mathsf{v}}(xy) &= e^{-N\left[H(z|x,y,\{u_i\})+H(x|\{u_i\})+H(y|\{u_i\})+H(\{u_i\})\right]} \\ &= e^{-N\left[H(x,y,z,\{u_i\})+I(x;y|\{u_i\})\right]}\end{aligned} \tag{9}$$

where $I(x; y|\{u_i\})$ is the conditional mutual information between $x$ and $y$ given $\{u_i\}$, $I(x; y|\{u_i\}) = H(x|\{u_i\}) + H(y|\{u_i\}) - H(x, y|\{u_i\}) - H(\{u_i\})$.

Likewise, the maximum likelihood $\mathcal{L}_{\mathsf{nv}}(xy)$ of a generalized non-v-structure of base $xy$ with upstream nodes $\{u_i\}_n$ and $z$ can be expressed as,

$$\begin{aligned}\mathcal{L}_{\mathsf{nv}}(xy) &= e^{-N\left[H(x|z,\{u_i\})+H(y|z,\{u_i\})+H(z,\{u_i\})\right]} \\ &= e^{-N\left[H(x,y,z,\{u_i\})+I(x;y|z,\{u_i\})\right]}\end{aligned} \tag{10}$$

where $I(x; y|z, \{u_i\}) = H(x|z, \{u_i\}) + H(y|z, \{u_i\}) - H(x, y|z, \{u_i\}) - H(z, \{u_i\})$ is the conditional mutual information between $x$ and $y$ given $z$ and $\{u_i\}$. Hence,

$$\frac{\mathcal{L}_{\mathsf{v}}(xy)}{\mathcal{L}_{\mathsf{nv}}(xy)} = e^{-NI(x;y;z|\{u_i\})} \tag{11}$$

where we introduced the conditional 3-point information, $I(x; y; z|\{u_i\}) = I(x; y|\{u_i\}) - I(x; y|z, \{u_i\})$.

Hence, a significantly negative conditional 3-point information, $I(x; y; z|\{u_i\}) < 0$, implies that a generalized v-structure is more likely than a generalized non-v-structure given the available observational data. Conversely, a significantly positive conditional 3-point information, $I(x; y; z|\{u_i\}) > 0$, implies that a generalized non-v-structure model is more likely than a generalized v-structure model.

Yet, as the conditional 3-point information, $I(x; y; z|\{u_i\})$, is in fact invariant upon permutations between $x, y$ and $z$, it cannot indicate how to orient embedded v-structures or non-v-structures over the $xyz$ triple, as already noted in the case of isolated v-structures and non-v-structures, above.

However, the most likely base ($xy$, $xz$ or $yz$) of the embedded v-structure or non-v-structure corresponds to the least correlated pair conditioned on $\{u_i\}$, *e.g.*, $I(x; y|\{u_i\}) = \min_{xyz}\big(I(s; t|\{u_i\})\big)$, as shown with the following likelihood ratios,

$$\frac{\mathcal{L}_{\mathsf{v}}(xy)}{\mathcal{L}_{\mathsf{v}}(st)} = \frac{\mathcal{L}_{\mathsf{nv}}(xy)}{\mathcal{L}_{\mathsf{nv}}(st)} = \frac{e^{-NI(x;y|\{u_i\})}}{e^{-NI(s;t|\{u_i\})}} \tag{12}$$

Note, in particular, that choosing the base with the lowest conditional mutual information, *e.g.*, $I(x; y|\{u_i\}) = \min_{xyz}\big(I(s; t|\{u_i\})\big)$, is consistent with the Data Processing Inequality expected for the generalized non-v-structure of Fig. 1f–h, $I(x; y) \leqslant \min\big(I(x; z, \{u_i\}), I(z, \{u_i\}; y)\big)$, as shown below for $I(x; y)$ and $I(x; z, \{u_i\})$, by subtracting $I(x; y; z|\{u_i\})$ on each side of the inequality $I(x; y|\{u_i\}) \leqslant I(x; z|\{u_i\})$, leading to,

$$\begin{aligned} I(x; y|z, \{u_i\}) &\leqslant I(x; z|\{u_i\}, y) \\ &\leqslant I(x; z|\{u_i\}, y) + I(x; \{u_i\}|y) \\ &\leqslant I(x; z, \{u_i\}|y) \\ I(x; y) &\leqslant I(x; z, \{u_i\}) \end{aligned} \tag{13}$$

where we have used the chain rule, $I(x; z, \{u_i\}|y) = I(x; z|\{u_i\}, y) + I(x; \{u_i\}|y)$, before adding $I(x; y; z, \{u_i\})$ on each side of the inequality. The corresponding inequality holds between $I(x; y)$ and $I(z, \{u_i\}; y)$, implying the Data Processing Inequality.

#### *Finite size corrections of maximum likelihood*

Maximum likelihood ratios, such as Eq. 2, suggest that $1/N$ sets the significance level of the maximum likelihood approach, as $H(\mathcal{G}, \mathcal{D}) - H(\mathcal{G}', \mathcal{D}) \gg 1/N$ should imply a significant improvement of the underlying model $\mathcal{G}'$ over $\mathcal{G}$. In practice, however, there are $\mathcal{O}(\log(N)/N)$ corrections coming from the proper normalization of maximum likelihoods (see Appendix),

$$\mathcal{L}_{\mathcal{G}} = \frac{e^{-N\sum_i H(x_i|\{\mathrm{Pa}_{x_i}\})}}{Z(\mathcal{G}, \mathcal{D})} \tag{14}$$

The model $\mathcal{G}$ can then be compared to the alternative model $\mathcal{G}_{\backslash x\to y}$ with one missing edge $x \to y$ using the maximum likelihood ratio,

$$\frac{\mathcal{L}_{\mathcal{G}_{\backslash x\to y}}}{\mathcal{L}_{\mathcal{G}}} = e^{-NI(x;y|\{\mathrm{Pa}_y\}_{\backslash x})}\frac{Z(\mathcal{G}, \mathcal{D})}{Z(\mathcal{G}_{\backslash x\to y}, \mathcal{D})} \tag{15}$$

where $I(x; y|\{\mathrm{Pa}_y\}_{\backslash x}) = H(y|\{\mathrm{Pa}_y\}_{\backslash x}) - H(y|\{\mathrm{Pa}_y\})$.

Then, following the rationale of constraint-based approaches, Eq. 15 can be reformulated by replacing the parent nodes $\{\mathrm{Pa}_y\}_{\backslash x}$ with an unknown separation set $\{u_i\}$ to be learnt simultaneously with the missing edge candidate $xy$,

$$\frac{\mathcal{L}_{\mathcal{G}_{\backslash xy|\{u_i\}}}}{\mathcal{L}_{\mathcal{G}}} = e^{-NI(x;y|\{u_i\})+k_{x;y|\{u_i\}}} \tag{16}$$

$$k_{x;y|\{u_i\}} = \log\left(Z(\mathcal{G},\mathcal{D})/Z(\mathcal{G}_{\backslash xy|\{u_i\}},\mathcal{D})\right) \tag{17}$$

where the factor $k_{x;y|\{u_i\}} > 0$ tends to limit the complexity of the models by favoring fewer edges. Namely, the condition, $I(x;y|\{u_i\}) < k_{x;y|\{u_i\}}/N$, implies that simpler models compatible with the structural independency, $x \perp\!\!\!\perp y|\{u_i\}$, are more likely than model $\mathcal{G}$, given the finite available dataset. This replaces the 'perfect' conditional independency condition, $I(x;y|\{u_i\}) = 0$, valid in the limit of an infinite dataset, $N \to \infty$. A common complexity criteria in model selection is the Bayesian Information Criteria (BIC) or Minimal Description Length (MDL) criteria [19, 20],

$$k^{\mathrm{MDL}}_{x;y|\{u_i\}} = \frac{1}{2}(r_x - 1)(r_y - 1)\prod_i r_{u_i}\log N \tag{18}$$

where $r_x, r_y$ and $r_{u_i}$ are the number of levels of the corresponding variables. The MDL complexity, Eq. 18, is simply related to the normalisation constant of the distribution reached in the asymptotic limit of a large dataset $N \to \infty$ (Laplace approximation). However, this limit distribution is only reached for very large datasets in practice.

Alternatively, the normalisation of the maximum likelihood can also be done over all possible datasets including the same number of data points to yield a (universal) Normalized Maximum Likelihood (NML) criteria [21, 22] and its decomposable [23, 24] and $xy$-symmetric version, $k^{\mathrm{NML}}_{x;y|\{u_i\}}$, defined in the Appendix.

Then, incrementing the separation set of $xy$ from $\{u_i\}$ to $\{u_i\} + z$ leads to the following likelihood ratio,

$$\frac{\mathcal{L}_{\mathcal{G}_{\backslash xy|\{u_i\},z}}}{\mathcal{L}_{\mathcal{G}_{\backslash xy|\{u_i\}}}} = e^{NI(x;y;z|\{u_i\})+k_{x;y;z|\{u_i\}}} \tag{19}$$

with $I(x;y;z|\{u_i\}) = I(x;y|\{u_i\}) - I(x;y|\{u_i\},z)$ and where we introduced a 3-point conditional complexity, $k_{x;y;z|\{u_i\}}$, defined similarly as the difference between the 2-point conditional complexities,

$$k_{x;y;z|\{u_i\}} = k_{x;y|\{u_i\},z} - k_{x;y|\{u_i\}} \tag{20}$$

However, unlike 3-point information, $I(x;y;z|\{u_i\})$, 3-point complexities are always positive, $k_{x;y;z|\{u_i\}} > 0$, provided that there are at least two levels for each implicated node $\ell \in x, y, z, \{u_i\}$, *i.e.* $r_\ell \geqslant 2$.

Hence, we can define the shifted 2-point and 3-point information in Eqs. 16 & 19 for finite datasets as,

$$I'(x;y|\{u_i\}) = I(x;y|\{u_i\}) - \frac{k_{x;y|\{u_i\}}}{N} \tag{21}$$

$$I'(x;y;z|\{u_i\}) = I(x;y;z|\{u_i\}) + \frac{k_{x;y;z|\{u_i\}}}{N} \tag{22}$$

This leads to the following maximum likelihood ratios equivalent to Eqs. 11 & 12 for v-structure over non-v-structure and between alternative bases,

$$\frac{\mathcal{L}_{\mathsf{v}}(xy)}{\mathcal{L}_{\mathsf{nv}}(xy)} = e^{-NI'(x;y;z|\{u_i\})} \tag{23}$$

$$\frac{\mathcal{L}_{\mathsf{v}}(xy)}{\mathcal{L}_{\mathsf{v}}(st)} = \frac{\mathcal{L}_{\mathsf{nv}}(xy)}{\mathcal{L}_{\mathsf{nv}}(st)} = \frac{e^{-NI'(x;y|\{u_i\})}}{e^{-NI'(s;t|\{u_i\})}} \tag{24}$$

Hence, given a finite dataset, a significantly negative conditional 3-point information, corresponding to $I'(x;y;z|\{u_i\}) < 0$, implies that a v-structure $x \to z \leftarrow y$ is more likely than a non-v-structure provided that the structural independency, $x \perp\!\!\!\perp y|\{u_i\}$, is also confidently established as, $I'(x;y|\{u_i\}) < 0$. By contrast, a significantly positive conditional 3-point information corresponds to $I'(x;y;z|\{u_i\}) > 0$ and implies that a non-v-structure model is more likely than a v-structure model, given the available observational data.

#### *Probability estimate of indirect contributions to mutual information*

The previous results enable us to estimate the probability of a node $z$ to contribute to the conditional mutual information $I(x;y|\{u_i\})$, by combining the probability, $P_{\mathsf{nv}}(xyz|\{u_i\})$, that the triple $xyz$ is a generalized non-v-structure conditioned on $\{u_i\}$ and the probability, $P_{\mathsf{b}}(xy|\{u_i\})$, that its base is $xy$, where,

$$P_{\mathsf{nv}}(xyz|\{u_i\}) = \frac{\mathcal{L}_{\mathsf{nv}}(xy)}{\mathcal{L}_{\mathsf{nv}}(xy) + \mathcal{L}_{\mathsf{v}}(xy)} \tag{25}$$

$$P_{\mathsf{b}}(xy|\{u_i\}) = \frac{\mathcal{L}_{\mathsf{nv}}(xy)}{\mathcal{L}_{\mathsf{nv}}(xy) + \mathcal{L}_{\mathsf{nv}}(xz) + \mathcal{L}_{\mathsf{nv}}(yz)} \tag{26}$$

that is, using Eqs. 23 & 24 including finite size corrections of the maximum likelihoods,

$$P_{\mathsf{nv}}(xyz|\{u_i\}) = \frac{1}{1 + e^{-NI'(x;y;z|\{u_i\})}} \tag{27}$$

$$P_{\mathsf{b}}(xy|\{u_i\}) = \frac{1}{1 + \frac{e^{-NI'(x;z|\{u_i\})}}{e^{-NI'(x;y|\{u_i\})}} + \frac{e^{-NI'(y;z|\{u_i\})}}{e^{-NI'(x;y|\{u_i\})}}} \tag{28}$$

Then, various alternatives to combine $P_{\mathsf{nv}}(xyz|\{u_i\})$ and $P_{\mathsf{b}}(xy|\{u_i\})$ exist to estimate the overall probability that the additional node $z$ indirectly contributes to $I(x;y|\{u_i\})$. One possibility is to choose the lower bound $S_{\mathsf{lb}}(z;xy|\{u_i\})$ of $P_{\mathsf{nv}}(xyz|\{u_i\})$ and $P_{\mathsf{b}}(xy|\{u_i\})$, since both conditions

need to be fulfilled to warrant that $z$ indeed contributes to $I(x;y|\{u_i\})$,

$$S_{\mathsf{lb}}(z;xy|\{u_i\}) = \min\left[P_{\mathsf{nv}}(xyz|\{u_i\}), P_{\mathsf{b}}(xy|\{u_i\})\right] \tag{29}$$

The pair of nodes $xy$ with the most likely contribution from a third node $z$ can then be ordered according to their rank $R(xy;z|\{u_i\})$ defined as,

$$R(xy;z|\{u_i\}) = \max_z \left(S_{\mathsf{lb}}(z;xy|\{u_i\})\right) \tag{30}$$

and $z$ can be iteratively added to the set of contributing nodes (*i.e.* $\{u_i\} \leftarrow \{u_i\} + z$) of the top link $xy = \mathrm{argmax}_{xy} R(xy;z|\{u_i\})$ to progressively recover the most significant indirect contributions to all pairwise mutual information in a causal graph, as outlined below.

#### *Robust inference of conditional independencies using the* 3off2 *scheme*

The previous results can be used to provide a robust inference method to identify conditional independencies and, hence, reconstruct the skeleton of underlying causal graphs from finite available observational data. The approach follows the spirit of constraint-based methods, such as the PC or IC algorithms, but recovers conditional independencies following an evolving ranking of the network edges, $R(xy;z|\{u_i\})$, defined in Eq. 30.

All in all, this amounts to perform a generic decomposition for each mutual information term, $I(x;y)$, by introducing a succession of node candidates, $u_1, u_2, \ldots, u_n$, that are likely to contribute to the overall mutual information between the pair $x$ and $y$, as,

$$\begin{aligned} I(x;y) &= I(x;y;u_1) + I(x;y|u_1) \\ &= I(x;y;u_1) + I(x;y;u_2|u_1) + \ldots \\ &\quad \ldots + I(x;y;u_n|\{u_i\}_{n-1}) + I(x;y|\{u_i\}_n) \end{aligned} \tag{31}$$

or equivalently between the shifted 2-point and 3-point information terms including finite size corrections (Eq. 22),

$$\begin{aligned} I'(x;y) &= I'(x;y;u_1) + I'(x;y;u_2|u_1) + \ldots \\ &\quad + I'(x;y;u_n|\{u_i\}_{n-1}) + I'(x;y|\{u_i\}_n) \end{aligned} \tag{32}$$

Hence, given a significant mutual information between $x$ and $y$, $I'(x;y) > 0$, we will search for possible structural independencies, *i.e.* $I'(x;y|\{u_i\}_n) < 0$, by iteratively "*taking off*" conditional 3-point information terms from the initial 2-point (mutual) information, $I'(x;y)$, as

$$\begin{aligned} I'(x;y|\{u_i\}_n) &= I'(x;y) - I'(x;y;u_1) - I'(x;y;u_2|u_1) \\ &\quad - \ldots - I'(x;y;u_n|\{u_i\}_{n-1}) \end{aligned} \tag{33}$$

and similarly with non-shifted 2-point and 3-point information,

$$\begin{aligned} I(x;y|\{u_i\}_n) &= I(x;y) - I(x;y;u_1) - I(x;y;u_2|u_1) \\ &\quad - \ldots - I(x;y;u_n|\{u_i\}_{n-1}) \end{aligned} \tag{34}$$

#### 3off2 **algorithm**

The 3off2 scheme can be used to devise a two-step algorithm (see Algorithm 2), inspired by constraint-based approaches, to first reconstruct network skeleton (Algorithm 2, step 1) before combining orientation and propagation of edges in a single step based on likelihood ratios (Algorithm 2, step 2).

#### *Reconstruction of network skeleton*

The 3off2 scheme will first be applied to iteratively remove edges with *maximum positive contributions*, $I'(x;y;u_k|\{u_i\}_{k-1}) > 0$, corresponding to the *most likely generalized non-v-structures* (Eq. 23), while *minimizing simultaneously* the remaining 2-point information, $I'(x;y|\{u_i\}_k)$ (Eq. 24), consistently with the data processing inequality. Such 3off2 scheme (Algorithm 2, step 1) will therefore progressively lower the conditional 2-point information terms, $I'(x;y) > \cdots > I'(x;y|\{u_i\}_{k-1}) > I'(x;y|\{u_i\}_k)$ and might ultimately result in the removal of the corresponding edge, $xy$, but only when a structural independency is actually found, *i.e.* $I'(x;y|\{u_i\}_n) < 0$, as in constraint-based algorithms for a given significance level $\alpha$. Yet, the skeleton obtained with the 3off2 scoring approach is expected to be more robust to finite observational data than the skeleton obtained with PC or IC algorithms, as the former results only from statistically significant 3-point contributions, $I'(x;y;u_k|\{u_i\}_{k-1}) > 0$, based on their quantitative 3off2 ranks, $R(xy;u_k|\{u_i\}_{k-1})$.

The best results on benchmark networks using these quantitative 3off2 ranks are obtained with the NML score (see Results and discussion Section below). The MDL score leads to equivalent results, as expected, in the limit of very large datasets (see Appendix). However, with smaller datasets, the most reliable results with the MDL score are obtained using *non-shifted* instead of shifted 2-point and 3-point information terms in the 3off2 rank of individual edges, Eq. 30. This is because the MDL complexity tends to underestimate the importance of edges between nodes with many levels (see Appendix). For finite datasets, it easily leads to spurious conditional independencies, $I'(x;y|\{ui\}) < 0$, when using shifted 2-point and 3-point information, Eq. 33, whereas using non-shifted information in the 3off2 ranks (Eq. 30) tends to limit the number of false negatives as early errors in $\{u_i\}$ can only increase $I(x;y|\{ui\}) \geqslant 0$, in the end, in Eq. 34.

#### *Orientation of network skeleton*

The skeleton and the separation sets resulting from the 3off2 iteration step (Algorithm 2, step 1) can then be used to orient the edges and to propagate orientations to the unshielded triples. However, while the constraint-based methods distinguish the v-structures orientation

step (Algorithm 1, step 2) from the propagation procedure (Algorithm 1, step 3), the 3off2 algorithm intertwines these two steps based on the respective likelihood scores of individual v-structures and non-v-structures (Algorithm 2, step 2).

As stated earlier, the magnitude and sign of the conditional 3-point information, $I(x;y;z|\{u_i\})$ (or equivalently the shifted 3-point information, Eq. 23), indicate if a non v-structure is more likely than a v-structure. Hence, all the unshielded triples can be ranked by the *absolute* value of their conditional 3-point information, that is, in decreasing order of their *likelihood* of being either a v-structure or a non-v-structure. As detailed in the step 2 of Algorithm 2, the most likely v-structure is used to set the first orientations, following $R_0$ orientation rule. The possible propagations are then performed, following $R_1$ propagation rule, starting from the unshielded triple having the most positive conditional 3-point information. The following most likely v-structure is considered when no further propagation is possible on unshielded triples with greater absolute 3-point information. If conflicting orientations arise (such as $a \to b \leftarrow c$ & $b \to c \leftarrow d$), the less likely v-structure and its possible propagations are ignored.

Note that we only implement the $R_0$ and $R_1$ propagation rules, which are applied in decreasing order of likelihood. In particular, we do not consider propagation rules $R_2$ and $R_3$ which are not associated to likelihood scores but enforce the hypothesis of acyclic constraint.

As for the 3off2 skeleton reconstruction, the orientation/propagation step of 3off2 allows for a robust discovery of orientations from finite observational data as it relies on a quantitative framework of likelihood ratios taken in decreasing order of their statistical significance. During this step, 3off2 recovers and propagates as many orientations as possible in an iterative procedure following the decreasing ranks of the unshielded triples based on the absolute value of their conditional 3-point information, $|I'(x;y;z|\{u_i\})|$.

**Algorithm 2:** 3off2 Network Reconstruction

**In:** finite observational dataset of size $N$;
complexity $k_{x;y|\{u_i\}}$
**Out:** (partially) oriented graph $\mathcal{G}$
**0. Initiation**
Start with complete undirected graph $\mathcal{G}$
**forall the** *links xy* **do**
  **if** $I(x;y) < k_{x;y|\emptyset}/N$ *i.e.* $I'(x;y) < 0$ **then**
    ***xy* link is** non-essential and **removed**
    **separation set** of $xy$: $\text{Sep}_{xy} = \emptyset$
  **else**
    find the **most contributing node** $z$ neighbor of $x$ or $y$ and **compute** 3off2 **rank**, $R(xy;z|\emptyset)$
  **end**
**end**
**1. Iteration**
**while** $\exists$ *xy link with* $R(xy;z|\{u_i\}) > 1/2$ **do**
  **for** *top link xy with highest rank* $R(xy;z|\{u_i\})$ **do**
    **expand contributing set** $\{u_i\} \leftarrow \{u_i\} + z$
    **if** $I(x;y|\{u_i\}) < k_{x;y|\{u_i\}}/N$ *i.e.* $I'(x;y|\{u_i\}) < 0$ **then**
      ***xy* link is** non-essential and **removed**
      **separation set** of $xy$: $\text{Sep}_{xy} = \{u_i\}$
    **else**
      find **next most contributing node** $z$ neighbor of $x$ or $y$ and **compute new** 3off2 **rank** of $xy$: $R(xy;z|\{u_i\})$
    **end**
    **sort the** 3off2 **rank list** $R(xy;z|\{u_i\})$
  **end**
**end**
**2. Orientation / Propagation**
Sort list of unshielded triples, $\mathcal{L}_c = \{\langle x,z,y\rangle_{x \not\sim y}\}$, in decreasing order of $|I'(x;y;z|\{u_i\})|$
**repeat**
  Take $\langle x,z,y\rangle_{x \not\sim y} \in \mathcal{L}_c$ with highest $|I'(x;y;z|\{u_i\})|$ on which $R_0$ or $R_1$ orientation rule can be applied
  **if** $I'(x;y;z|\{u_i\}) < 0$ **then**
    if $\langle x,z,y\rangle_{x \not\sim y}$ has no diverging orientation, apply $R_0$: $\{x - * z * - y$ & $x \not\sim y$ & $z \notin \text{Sep}_{xy}\} \Rightarrow \{x \to z \leftarrow y\}$
  **else**
    if $\langle x,z,y\rangle_{x \not\sim y}$ has one converging orientation, apply $R_1$: $\{x \to z - y$ & $x \not\sim y\} \Rightarrow \{z \to y\}$
  **end**
  Apply new orientation(s) to all other $\langle x',z',y'\rangle_{x' \not\sim y'} \in \mathcal{L}_c$
**until** *no additional orientation can be obtained*;

## Results and discussion

### Tests on benchmark graphs

We have tested the 3off2 network reconstruction approach to learn benchmark causal graphs containing 20 to 70 nodes, Figs. 2, 3, 4, 5 and 6. The results are evaluated against other methods in terms of Precision (or positive predictive value), $Prec = TP/(TP + FP)$, Recall or Sensitivity (true positive rate), $Rec = TP/(TP + FN)$, as well as F-score $= 2 \times Prec \times Rec/(Prec + Rec)$ for increasing sample size $N$=10 to 50,000 data points.

We also define additional Precision, Recall and F-scores taking into account the edge orientations of the predicted networks against the corresponding CPDAG of the benchmark networks. This amounts to label as false positives, all true positive edges of the skeleton

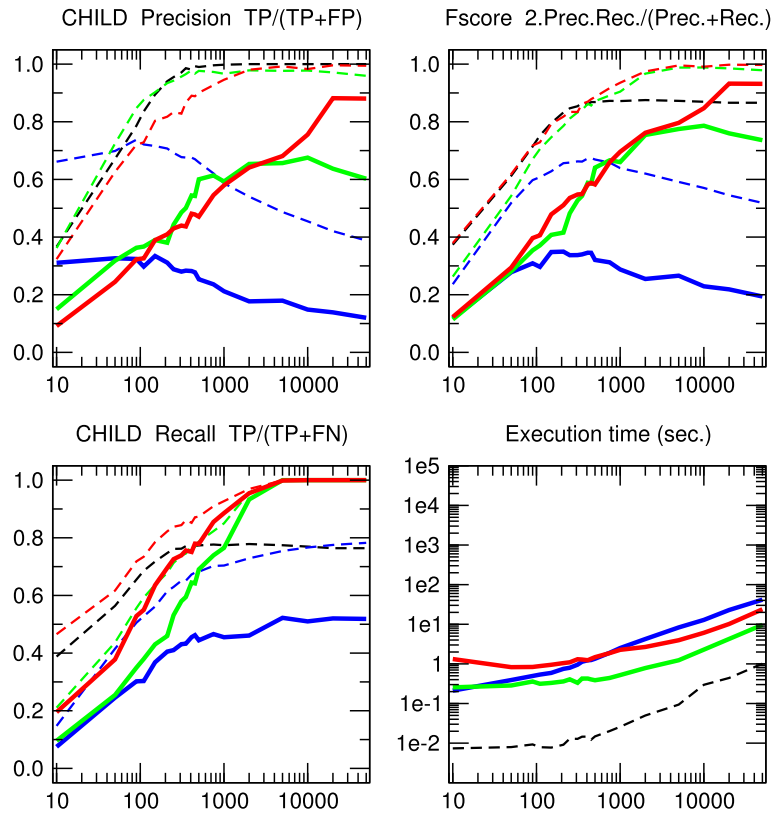



**Fig. 2** CHILD network. [20 nodes, 25 links, 230 parameters, Average degree 2.5, Maximum in-degree 2]. Precision, Recall and F-score for skeletons (*dashed lines*) and CPDAGs (*solid lines*). The results are given for Aracne (*black*), PC (*blue*), Bayesian Hill-Climbing (*green*) and 3off2 (*red*)

with different orientation/non-orientation status as the CPDAG reference, $TP_{\text{misorient}}$, leading to the orientation-dependent definitions $TP' = TP - TP_{\text{misorient}}$ and $FP' = FP + TP_{\text{misorient}}$ with the corresponding CPDAG Precision, Recall and F-scores taking into account edge orientations.

The alternative inference methods used for comparison with 3off2 are the PC algorithm [12] implemented in the `pcalg` package [25, 26] and Bayesian inference using the hill-climbing heuristics implemented in the `bnlearn` package [27]. In addition, we also compare the skeleton of 3off2 to the unoriented output of Aracne [28], an information-based inference approach, which iteratively prunes links with the weakest mutual information based on the Data Processing Inequality. We have used the Aracne implementation of the `minet` package [29]. For each sample size, 3off2, Aracne, PC and the Bayesian inference methods have been tested on 50 replicates. Figures 2, 3, 4, 5 and 6 give the average results over these multiple replicates when comparing the CPDAG (solid lines) of the reconstructed network (or its skeleton, dashed lined) to the CPDAG (or the skeleton) of the benchmark network.

For each method, the plots presented in Figs. 2, 3, 4, 5 and 6 are those obtained for the parameters that give overall the best results over the five reconstructed benchmark networks (see Additional file 1, Figures S1-S20). In particular, we used the *stable* implementation of the PC algorithm, as well as the *majority rule* for the orientation and propagation steps [14]. PC's results are shown on Figs. 2, 3, 4, 5 and 6 for $\alpha = 0.1$. Decreasing $\alpha$ tends to improve the skeleton Precision at the expense of the skeleton Recall, leading in fact to worse skeleton F-scores for finite datasets, *e.g.* $N \leqslant 1000$ (see Additional file 1, Figures S1-S5). The same trend is observed for CPDAG F-scores taking into account edge orientations, with best CPDAG scores at small sample sizes, obtained for larger $\alpha$, *e.g.* $N \leqslant 1000$. Aracne threshold parameters for minimum difference in mutual information is set to $\epsilon = 0$, as small positive values typically worsen F-scores (see Additional file 1, Figures S6-S10). Bayesian inference are obtained using BIC/MDL scores and hill-climbing heuristics with 100 random restarts [9] (see Additional file 1, Figures S11-S15). Finally, the best 3off2 network reconstructions are obtained using NML scores with shifted 2-point and 3-point information terms in the rank of individual edges,

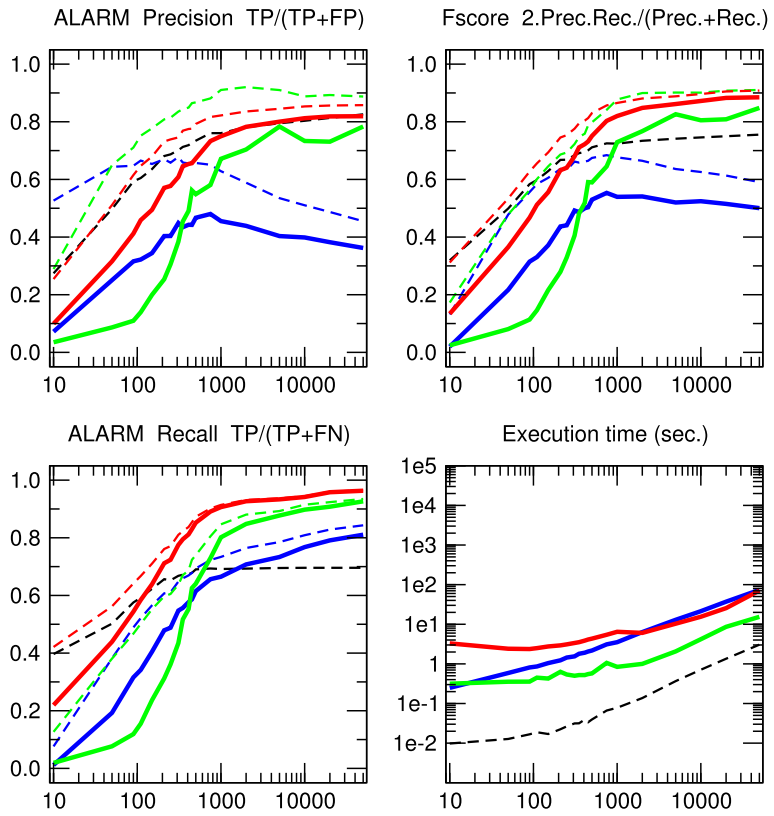



**Fig. 3** ALARM network. [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall and F-score for skeletons (*dashed lines*) and CPDAGs (*solid lines*). The results are given for Aracne (*black*), PC (*blue*), Bayesian Hill-Climbing (*green*) and 3off2 (*red*)

see Methods. Using MDL scores, instead, leads to equivalent results, as expected, in the limit of very large datasets (see Appendix). However, with smaller datasets, the most reliable results with MDL scores are obtained using *non-shifted* instead of shifted 2-point and 3-point information terms in the 3off2 rank of individual edges, as discussed in Methods (see Additional file 1, Figures S16-S20).

All in all, we found that the 3off2 inference approach typically reaches better or equivalent F-scores for all dataset sizes as compared to all other tested methods, *i.e.* Aracne, PC and Bayesian inference, as well as the Max-Min Hill-Climbing (MMHC) hybrid method [30] (see Additional file 1, Figures S21-S25). This is clearly observed both on the skeletons (Figs. 2, 3, 4, 5 and 6 dashed lines) and even more clearly when taking the predictions of orientations into account (Figures 2, 3, 4, 5 and 6 solid lines).

### Applications to the hematopoiesis regulation network

The reconstruction or reverse-engineering of real regulatory networks from actual expression data has already been performed on a number of biological systems (see *e.g.* [28, 31–33]). Here, we apply the 3off2 approach on a real biological dataset related to hematopoiesis. Transcription factors play a central role in hematopoiesis, from which derive the blood cell lineages. As suggested in previous studies, changes in the regulatory interactions among transcription factors [34] or their overexpression [35] might be involved in the development of T-acute lymphoblastic leukaemia (T-ALL). The key role of the hematopoiesis and the potentially serious consequences of its disregulations emphasize the need to accurately establish the complex interactions between the transcription factors involved in this critical biological process.

The dataset we have used for this analysis [36] consists of the single cell expressions of 18 transcription factors, known for their role in hematopoiesis. Five hundred ninety seven single cells representing 5 different types of hematopoietic progenitors have been included in the analysis ($N = 597$). We reconstructed the corresponding network with the 3off2 inference method, Fig. 7, and four other available approaches, namely, PC [12] implemented in the `pcalg` package [25, 26], Bayesian inference using hill-climbing heuristics as well as the Max-Min Hill-Climbing (MMHC) hybrid method [30], both implemented in the `bnlearn` package [27], and, finally, Aracne [28] implemented in the `minet` package [29] (Table 1 and Additional file 1: Table S1).

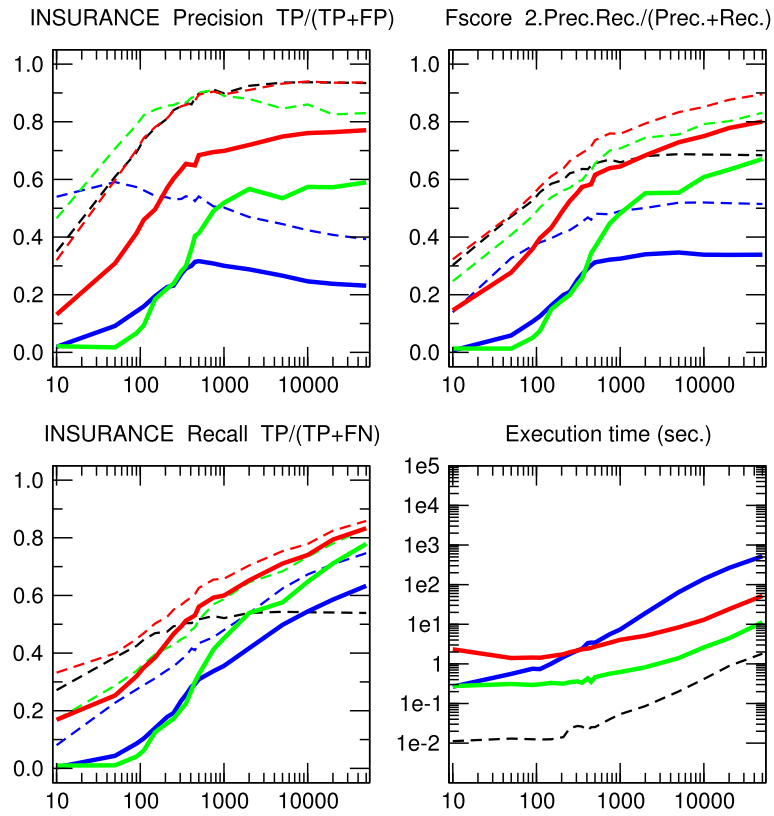



**Fig. 4** INSURANCE network. [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall and F-score for skeletons (*dashed lines*) and CPDAGs (*solid lines*). The results are given for Aracne (*black*), PC (*blue*), Bayesian Hill-Climbing (*green*) and 3off2 (*red*)

3off2 uncovers all 11 interactions for which specific experimental evidence has been reported in the literature (Fig. 7, red links: known activations; blue links: known repressions) as well as 30 additional links (Fig. 7, grey links: unknown regulatory interactions). By contrast, randomization of the actual data across samples for each TF leads to only 5.25 spurious interactions on average between the 18 TFs, instead of the 41 inferred edges from the actual data, and 1.62 spurious interactions on average, instead of the 16 interactions predicted among the 10 TFs involved in known regulatory interactions, Fig. 7. This suggests that around 10–13 % of the predicted edges might be spurious, due to inevitable sampling noise in the finite dataset. In particular, the 3off2 inference approach successfully recovers the relationships of the regulatory triad between *Gata2*, *Gfi1b* and *Gfi1* as described in [36] and reports correct orientations for the edges involving *Gata2* (*Gfi1b* and *Gfi1* crossregulate in fact one another [36], Table 1). The network reconstructed by 3off2 also correctly infers the regulations of *PU.1* by *Gfi1* [37], *Gfi1* by *Lyl1* [38], *Meis1* by *Ldb1* [39], and the regulations of *Lyl1* by *Ldb1* [39] and *Erg* [40]. Finally, the interactions (*Gata2*−*SCL*) [40], (*Gfi1b*−*Meis1*) [41] and (*Gata1*−*Gata2*) [42] are correctly inferred, however, with opposite directions as reported in the literature. Yet, overall 3off2 outperforms most of the other methods tested for the reconstruction of the hematopoietic regulatory subnetwork (Table 1 and Additional file 1: Table S1). Only the Bayesian hill-climbing method using a BDe score leads to comparable results by retrieving 10 out of 11 interactions and correctly orienting 8 of them. These encouraging results from the 3off2 reconstruction method on experimentally proven regulatory interactions (red edges in Fig. 7) could motivate further investigations on novel regulatory interactions awaiting to be tested for their possible role in hematopoiesis (*e.g.* grey edges in Fig. 7).

## Conclusions

In this paper, we propose to improve constraint-based network reconstruction methods by identifying structural independencies through a robust quantitative score-based scheme limiting the accumulation of early FN errors and subsequent FP compensatory errors. In brief, 3off2 relies on information theoretic scores to progressively uncover the best supported conditional independencies, by iteratively "taking off" the most likely indirect contributions of conditional 3-point information from every 2-point (mutual) information of the causal graph.

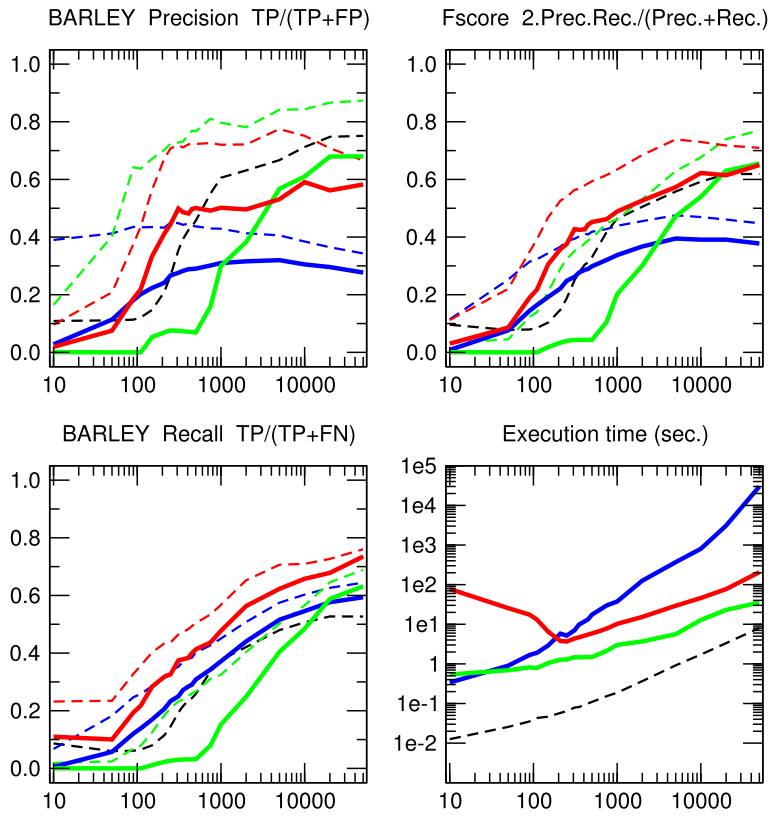



**Fig. 5** BARLEY network. [48 nodes, 84 links, 114,005 parameters, Average degree 3.5, Maximum in-degree 4]. Precision, Recall and F-score for skeletons (*dashed lines*) and CPDAGs (*solidlines*). The results are given for Aracne (*black*), PC (*blue*), Bayesian Hill-Climbing (*green*) and 3off2 (*red*)

Earlier hybrid methods have also attempted to improve network reconstruction by combining the concepts of constraint-based approaches with the robustness of Bayesian scores [30, 43–45]. In particular [43], have proposed to exploit an intrinsic weakness of the PC algorithm, its sensitivity to the order in which conditional independencies are tested on finite data, to rank these different order-dependent PC predictions with Bayesian scores. More recently [30], have also combined constraint-based and Bayesian approaches by first identifying both parents and children of each node of the underlying graphical model and then performing a greedy Bayesian hill-climbing search restricted to the identified parents and children of each node. This Max-Min Hill-Climbing (MMHC) approach tends to have a high precision in terms of skeleton but a more limited sensibility, leading overall to lower skeleton and CPDAG F-scores than 3off2 and Bayesian hill climbing methods on the same benchmark networks, Figures S21-S25. Interestingly, however, the MMHC approach is among the fastest network reconstruction approaches, Figure S26, allowing for scalability to large network sizes [30].

The 3off2 algorithm is expected to run in polynomial time on *typical* sparse causal networks with low in-degree, just like constraint-based algorithms. However, in practice and despite the additional computation of conditional 2-point and 3-point information terms, we found that the 3off2 algorithm runs typically faster than constraint-based algorithms for large enough samples, by avoiding the cascading accumulation of errors that inflate the combinatorial search of conditional independencies in traditional constraint-based approaches. Instead, we found that 3off2 running time displays a similar trend as Bayesian hill-climbing heuristic methods, Figs. 2, 3, 4, 5 and 6.

All in all, the main computational bottleneck of the present 3off2 scheme pertains to the identification of the *best* contributing nodes at each iteration. In the future, it could be interesting to investigate whether a more stochastic version of this 3off2 method, based on choosing *one* significant conditional 3-point information instead of the best one, might simultaneously accelerate the network reconstruction and circumvent possible locally trapped suboptimal predictions through stochastic resampling.

Finally, another perspective for practical applications will be to include the possibility of latent variables and bidirected edges in reconstructed networks.

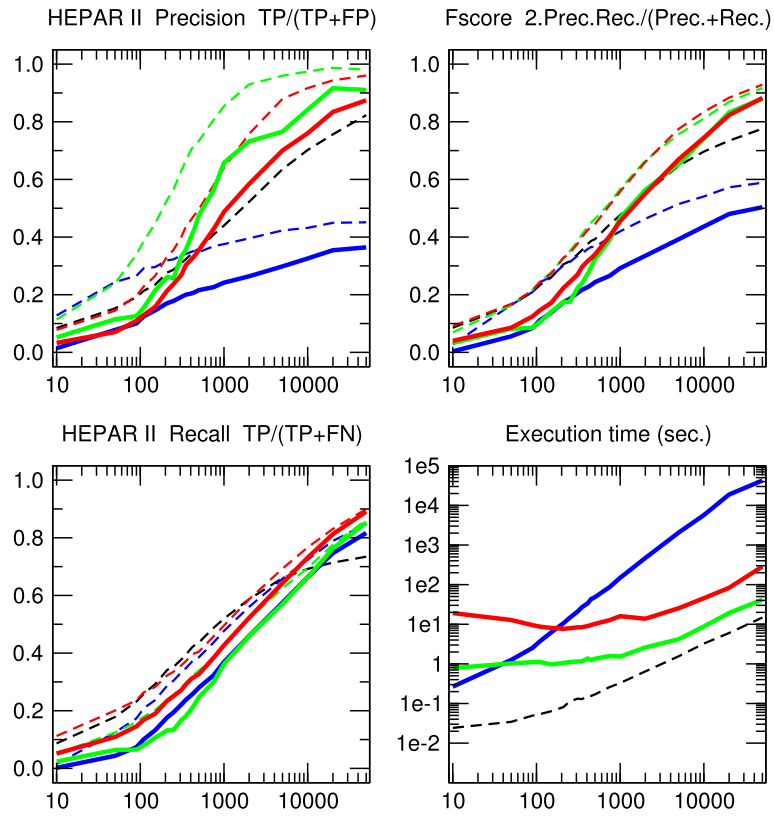



**Fig. 6** HEPAR II network. [70 nodes, 123 links, 1,453 parameters, Average degree 3.51, Maximum in-degree 6]. Precision, Recall and F-score for skeletons (*dashed lines*) and CPDAGs (*solid lines*). The results are given for Aracne (*black*), PC (*blue*), Bayesian Hill-Climbing (*green*) and 3off2 (*red*)

## Appendix

### Complexity of graphical models

The complexity $k_{\mathcal{G},\mathcal{D}}$ of a graphical model is related to the normalization constant $Z(\mathcal{G},\mathcal{D})$ of its maximum likelihood as $k_{\mathcal{G},\mathcal{D}} = \log Z(\mathcal{G},\mathcal{D})$,

$$\mathcal{L}_{\mathcal{G}} = \frac{e^{-NH(\mathcal{G},\mathcal{D})}}{Z(\mathcal{G},\mathcal{D})} = e^{-NH(\mathcal{G},\mathcal{D})-k_{\mathcal{G},\mathcal{D}}} \tag{35}$$

For Bayesian networks with decomposable entropy, *i.e.* $H(\mathcal{G},\mathcal{D}) = \sum_i H(x_i|\{\mathrm{Pa}_{x_i}\})$, it is convenient to use decomposable complexities, $k_{\mathcal{G},\mathcal{D}} = \sum_i k_{x_i|\{\mathrm{Pa}_{x_i}\}}$,

$$\mathcal{L}_{\mathcal{G}} = e^{-N\sum_i H(x_i|\{\mathrm{Pa}_{x_i}\})-\sum_i k_{x_i|\{\mathrm{Pa}_{x_i}\}}} \tag{36}$$

such that the comparison between alternative models $\mathcal{G}$ and $\mathcal{G}_{\backslash x\to y}$ (*i.e.* $\mathcal{G}$ with one missing edge $x \to y$) leads to a simple local increment of the score,

$$\frac{\mathcal{L}_{\mathcal{G}_{\backslash x\to y}}}{\mathcal{L}_{\mathcal{G}}} = e^{-NI(x;y|\{\mathrm{Pa}_y\}_{\backslash x})+\Delta k_{y|\{\mathrm{Pa}_y\}_{\backslash x}}} \tag{37}$$

$$I(x;y|\{\mathrm{Pa}_y\}_{\backslash x}) = H(y|\{\mathrm{Pa}_y\}_{\backslash x}) - H(y|\{\mathrm{Pa}_y\}) \geqslant 0 \tag{38}$$

$$\Delta k_{y|\{\mathrm{Pa}_y\}_{\backslash x}} = k_{y|\{\mathrm{Pa}_y\}} - k_{y|\{\mathrm{Pa}_y\}_{\backslash x}} \geqslant 0 \tag{39}$$

A common complexity criteria in model selection is the Bayesian Information Criteria (BIC) or Minimal Description Length (MDL) criteria [19, 20],

$$k^{\mathrm{MDL}}_{y|\{\mathrm{Pa}_y\}} = \frac{1}{2}(r_y - 1)\prod_j^{\mathrm{Pa}_y} r_j \ \log N \tag{40}$$

$$\Delta k^{\mathrm{MDL}}_{y|\{\mathrm{Pa}_y\}_{\backslash x}} = \frac{1}{2}(r_x - 1)(r_y - 1)\prod_j^{\mathrm{Pa}_{y_{\backslash x}}} r_j \ \log N \tag{41}$$

where $r_x, r_y$ and $r_j$ are the number of levels of each variable, $x$, $y$ and $j$. The MDL complexity, Eq. 40, is simply related to the normalisation constant reached in the asymptotic limit of a large dataset $N \to \infty$ (Laplace approximation). The MDL complexity can also be derived from the Stirling approximation on the Bayesian measure [46, 47]. Yet, in practice, this limit distribution is only reached for very large datasets, as some of the least-likely $(r_y - 1)\prod_j r_j$ combinations of states of variables are in fact rarely (if ever) sampled in typical finite datasets. As a result, the MDL complexity criteria tends to underestimate the relevance of edges connecting variables with many levels, $r_i$, leading to the removal of false negative edges.

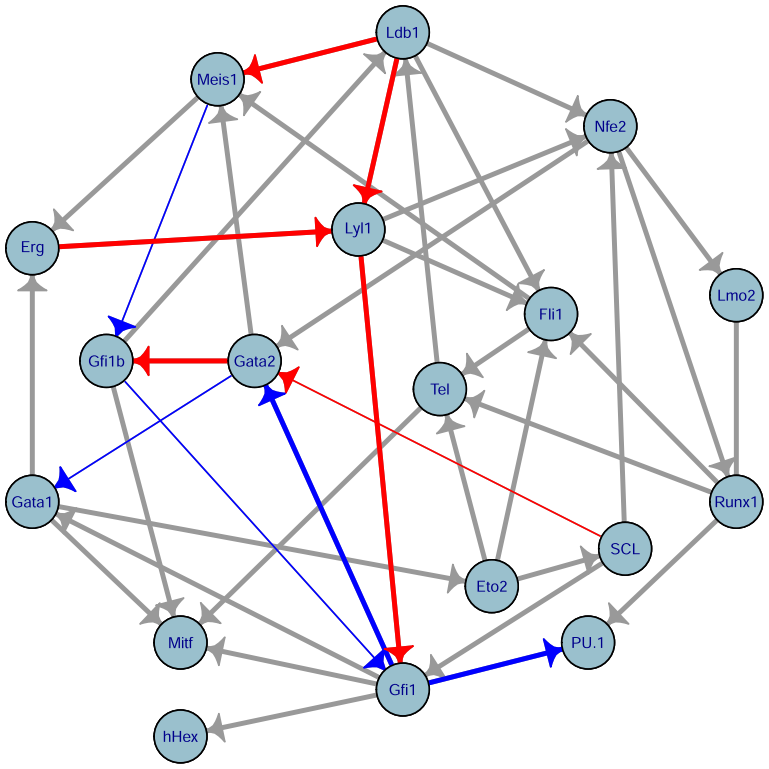



**Fig. 7** Hematopoietic subnetwork reconstructed by 3off2. The dataset [36] concerns 18 transcription factors, 597 single cells, 5 different hematopoietic progenitor types. Red and blue edges correspond to experimentally proven activations and repressions, respectively as reported in the literature (Table 1), while grey links indicate regulatory interactions for which no clear evidence has been established so far. Thinner arrows underline **3off2** misorientations

To avoid such biases with finite datasets, the normalisation of the maximum likelihood can be done over all possible datasets with the same number $N$ of data points. This corresponds to the (universal) Normalized Maximum Likelihood (NML) criteria [21–24],

$$\mathcal{L}_{\mathcal{G}} = \frac{e^{-NH(\mathcal{G},\mathcal{D})}}{\sum_{|\mathcal{D}'|=N} e^{-NH(\mathcal{G},\mathcal{D}')}} = e^{-NH(\mathcal{G},\mathcal{D})-k^{\mathsf{NML}}_{\mathcal{G},\mathcal{D}}} \tag{42}$$

We introduce here the factorized version of the NML criteria [23, 24] which corresponds to a decomposable NML score, $k^{\mathsf{NML}}_{\mathcal{G},\mathcal{D}} = \sum_{x_i} k^{\mathsf{NML}}_{x_i|\{\mathrm{Pa}_{x_i}\}}$, defined as,

$$k^{\mathsf{NML}}_{y|\{\mathrm{Pa}_y\}} = \sum_{j}^{q_y} \log \mathcal{C}^{r_y}_{N_{yj}} \tag{43}$$

$$\Delta k^{\mathsf{NML}}_{y|\{\mathrm{Pa}_y\}\setminus x} = \sum_{j}^{q_y} \log \mathcal{C}^{r_y}_{N_{yj}} - \sum_{j'}^{q_y/r_x} \log \mathcal{C}^{r_y}_{N_{yj'}} \tag{44}$$

where $N_{yj}$ is the number of data points corresponding to the $j$th state of the parents of $y$, $\{\mathrm{Pa}_y\}$, and $N_{yj'}$ the number of data points corresponding to the $j'$th state of the parents of $y$, excluding $x$, $\{\mathrm{Pa}_y\}\setminus x$. Hence, the factorized NML score for each node $x_i$ corresponds to a separate normalisation for each state $j = 1, \ldots, q_i$ of its parents and involving exactly $N_{ij}$ data points of the finite dataset,

$$\mathcal{L}_{\mathcal{G}} = e^{-N\sum_i H(x_i|\{\mathrm{Pa}_{x_i}\})-\sum_i\sum_j^{q_i}\log \mathcal{C}^{r_i}_{N_{ij}}} \tag{45}$$

$$= e^{N\sum_i\sum_j^{q_i}\sum_k^{r_i}\frac{N_{ijk}}{N}\log\left(\frac{N_{ijk}}{N_{ij}}\right)-\sum_i\sum_j^{q_i}\log \mathcal{C}^{r_i}_{N_{ij}}} \tag{46}$$

$$= \prod_i \prod_j^{q_i} \frac{\prod_k^{r_i}\left(\frac{N_{ijk}}{N_{ij}}\right)^{N_{ijk}}}{\mathcal{C}^{r_i}_{N_{ij}}} \tag{47}$$

where $N_{ijk}$ corresponds to the number of data points for which the $i$th node is in its $k$th state and its parents in their $j$th state, with $N_{ij} = \sum_k^{r_i} N_{ijk}$. The universal normalization constant $\mathcal{C}^r_n$ is then obtained by averaging over all possible partitions of the $n$ data points into a maximum of $r$ subsets, $\ell_1 + \ell_2 + \cdots + \ell_r = n$ with $\ell_k \geqslant 0$,

$$\mathcal{C}^r_n = \sum_{\ell_1+\ell_2+\cdots+\ell_r=n} \frac{n!}{\ell_1!\,\ell_2!\cdots\ell_r!}\prod_{k=1}^{r}\left(\frac{\ell_k}{n}\right)^{\ell_k} \tag{48}$$

which can in fact be computed in linear-time using the following recursion [23],

$$\mathcal{C}^r_n = \mathcal{C}^{r-1}_n + \frac{n}{r-2}\mathcal{C}^{r-2}_n \tag{49}$$

**Table 1** Interactions reconstructed by **3off2** and alternative methods for a subnetwork of hematopoiesis regulation. → indicates a successfully recovered interaction including its direction as reported in the literature (see References). ↛ corresponds to a successfully recovered interaction, however, with an opposite direction as reported in the literature. ≠ stipulates that no direct regulatory interaction has been inferred, while — corresponds to an undirected link. Note in particular that Aracne does not infer edge direction. See Additional file 1: Table S1 for supplementary statistics

| 11 known Regulatory interactions | References | 3off2 *NML* | PC $\alpha = 10^{-1}$ | PC $\alpha = 10^{-2}$ | MMHC *BDe* | MMHC *BIC* | Bayes hc *BDe* | Bayes hc *BIC* | Aracne $\epsilon = 0$ |
|---|---|---|---|---|---|---|---|---|---|
| Gata2 → Gfi1b | [36] | → | ↛ | — | ≠ | ≠ | → | ≠ | ≠ |
| Gfi1 → Gata2 | [36] | → | → | — | → | ↛ | → | ↛ | — |
| Gfi1b ⇆ Gfi1 | [36] | ↛ | ↛ | — | ↛ | ↛ | ↛ | ↛ | — |
| Gfi1 → PU.1 | [37] | → | → | ≠ | ≠ | ≠ | → | → | — |
| Lyl1 → Gfi1 | [38] | → | ↛ | ≠ | ≠ | ≠ | → | ↛ | — |
| Ldb1 → Meis1 | [39] | → | ≠ | ≠ | ≠ | ≠ | ↛ | ≠ | ≠ |
| Ldb1 → Lyl1 | [39] | → | ≠ | ≠ | ≠ | ≠ | ≠ | ≠ | ≠ |
| Erg → Lyl1 | [40] | → | ↛ | — | → | → | → | ↛ | — |
| Gata2 → Scl | [40] | ↛ | → | — | → | → | → | → | — |
| Gfi1b → Meis1 | [41] | ↛ | ↛ | — | → | → | → | → | — |
| Gata1 → Gata2 | [42] | ↛ | ↛ | — | → | → | → | → | — |
| Correct edges (out of **11**) | (→/↛/—) | **11** | **9** | **7** | **6** | **6** | **10** | **8** | **8** |
| - Correct orientations | (→) | 7 | 3 | 0 | 5 | 4 | 8 | 4 | 0 |
| - Mis/non-orientations | (↛/ —) | 4 | 6 | 7 | 1 | 2 | 2 | 4 | 8 |
| Missing links | (≠) | **0** | **2** | **4** | **5** | **5** | **1** | **3** | **3** |

with $\mathcal{C}_0^r = 1$ for all $r$, $\mathcal{C}_n^1 = 1$ for all $n$ and applying the general formula Eq. 48 for $r = 2$,

$$\mathcal{C}_n^2 = \sum_{h=0}^{n} \binom{n}{h} \left(\frac{h}{n}\right)^h \left(\frac{n-h}{n}\right)^{n-h} \tag{50}$$

or its Szpankowski approximation for large $n$ (needed for $n > 1000$ in practice) [48–50],

$$\mathcal{C}_n^2 = \sqrt{\frac{n\pi}{2}} \left(1 + \frac{2}{3}\sqrt{\frac{2}{n\pi}} + \frac{1}{12n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right) \tag{51}$$

$$\simeq \sqrt{\frac{n\pi}{2}} \exp\left(\sqrt{\frac{8}{9n\pi}} + \frac{3\pi - 16}{36n\pi}\right) \tag{52}$$

Then, following the rationale of constraint-based approaches, we can reformulate the likelihood ratio of Eq. 37 by replacing the parent nodes $\{\mathrm{Pa}_y\}_{\setminus x}$ in the conditional mutual information, $I(x;y|\{\mathrm{Pa}_y\}_{\setminus x})$, with an unknown separation set $\{u_i\}$ to be learnt simultaneously with the missing edge candidate $xy$,

$$\frac{\mathcal{L}_{\mathcal{G}_{\setminus xy|\{u_i\}}}}{\mathcal{L}_{\mathcal{G}}} = e^{-NI(x;y|\{u_i\}) + k_{x;y|\{u_i\}}} \tag{53}$$

where we have also transformed the asymmetric parent-dependent complexity difference, $\Delta k_{y|\{\mathrm{Pa}_y\}_{\setminus x}}$, into a $\{u_i\}$-dependent complexity term, $k_{x;y|\{u_i\}}$, with the same $xy$-symmetry as $I(x;y|\{u_i\})$,

$$k^{\mathrm{MDL}}_{x;y|\{u_i\}} = \frac{1}{2}(r_x - 1)(r_y - 1)\prod_i r_{u_i} \log N \tag{54}$$

$$k^{\mathrm{NML}}_{x;y|\{u_i\}} = \frac{1}{2}\sum_{j'}^{\{u_i\}} \Big( \sum_{k_x}^{r_x} \log \mathcal{C}^{r_y}_{N_{k_x j'}} - \log \mathcal{C}^{r_y}_{N_{j'}} + \sum_{k_y}^{r_y} \log \mathcal{C}^{r_x}_{N_{k_y j'}} - \log \mathcal{C}^{r_x}_{N_{j'}} \Big) \tag{55}$$

Note, in particular, that the MDL complexity term in Eq. 54 is readily obtained from Eq. 41 due to the Markov equivalence of the MDL score, corresponding to its $xy$-symmetry whenever $\{\mathrm{Pa}_y\}_{\setminus x} = \{\mathrm{Pa}_x\}_{\setminus y}$. By contrast, the factorized NML score, Eq. 43, is not a Markov-equivalent score (although its non-factorized version, Eq. 42, is Markov equivalent by definition). To circumvent this non-equivalence of factorized NML score, we propose to recover the expected $xy$-symmetry of $k^{\mathrm{NML}}_{x;y|\{u_i\}}$ through the simple $xy$-symmetrization of Eq. 44, leading to Eq. 55.

## Additional file

**Additional file 1: Complementary evaluations for the 3off2 inference approach and comparisons with alternative reconstruction methods and parameters values.** In this additional file, the results of the **3off2** inference approach are evaluated against other methods in terms of Precision (or positive predictive value), $Prec = TP/(TP + FP)$, Recall or Sensitivity (true positive rate), $Rec = TP/(TP + FN)$, as well as F-score $= 2 \times Prec \times Rec/(Prec + Rec)$ and execution time when comparing the CPDAG of the reconstructed network (or its skeleton) to the CPDAG (or the skeleton) of the benchmark network. The alternative methods are the PC algorithm, the Bayesian inference method using the hill-climbing heuristics, the Max-Min Hill-Climbing (MMHC) hybrid method and the Aracne inference approach. (PDF 528 KB)

**Competing interests**
The authors declare that they have no competing interests.

**Authors' contributions**
SA, LV and HI conceived and performed the research. SA, LV and HI wrote the manuscript. All authors read and approved the final manuscript.

**Acknowledgements**
S.A. acknowledges a PhD fellowship from the Ministry of Higher Education and Research and support from Fondation ARC pour la recherche sur le cancer. L.V. acknowledges a PhD fellowship from the Région Ile-de-France (DIM Institut des Systèmes Complexes) and H.I. acknowledges funding from CNRS, Institut Curie, Foundation Pierre-Gilles de Gennes and Région Ile-de-France.

**Publication costs**
Publication costs for this article were funded by the Région Ile-de-France.

**Declarations**
This article has been published as part of BMC Bioinformatics Volume 17 Supplement 2, 2016: Bringing Maths to Life (BMTL). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements.

Published: 20 January 2016

**References**

1. Cooper GF, Herskovits E. A bayesian method for the induction of probabilistic networks from data. Mach Learn. 1992;9(4):309–47.
2. Heckerman D, Geiger D, Chickering DM. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Mach Learn. 1995;20(3):197–243. Available as Technical Report MSR-TR-94-09.
3. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search, 2nd edn. Cambridge, MA: MIT press; 2000.
4. Pearl J. Causality: Models, Reasoning and Inference, 2nd edn: Cambridge University Press; 2009.
5. Chickering DM. Learning equivalence classes of bayesian-network structures. J Mach Learn Res. 2002;2:445–98.
6. Friedman N, Koller D. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. Mach Learn. 2003;50(1–2):95–125.
7. Koivisto M, Sood K. Exact bayesian structure discovery in bayesian networks. J Mach Learn Res. 2004;5:549–73.
8. Silander T, Myllymaki P. A simple approach for finding the globally optimal bayesian network structure. In: Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06). Arlington, Virginia: AUAI Press; 2006. p. 445–52.
9. Chickering DM, Geiger D, Heckerman D. Learning Bayesian networks: Search methods and experimental results. In: Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics; 1995. p. 112–28.
10. Bouckaert RR. Properties of bayesian belief network learning algorithms. In: Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence. UAI'94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1994. p. 102–9.
11. Friedman N, Nachman I, Pe'er D. Learning bayesian network structure from massive datasets: The "sparse candidate"; algorithm. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. UAI'99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1999. p. 206–15.
12. Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. Soc Sci Comput Rev. 1991;9:62–72.
13. Pearl J, Verma T. A theory of inferred causation. In: In Knowledge Representation and Reasoning: Proc. of the Second Int. Conf. San Mateo, CA: Morgan Kaufmann; 1991. p. 441–52.
14. Colombo D, Maathuis MH. Order-independent constraint-based causal structure learning. J Mach Learn Res. 2014;15:3741–782.
15. Meek C. Causal inference and causal explanation with background knowledge. In: Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, QU. San Francisco, CA: Morgan Kaufmann; 1995. p. 403–18.
16. Sanov IN. On the probability of large deviations of random variables. Mat Sbornik. 1957;42:11–44.
17. McGill WJ. Multivariate information transmission. Trans IRE Prof Group on Inf Theory (TIT). 1954;4:93–111.
18. Han TS. Multiple mutual informations and multiple interactions in frequency data. Inf Control. 1980;46(1):26–45.
19. Rissanen J. Modeling by shortest data description. Automatica. 1978;14:465–71.
20. Hansen MH, Yu B. Model selection and the principle of minimum description length. J Am Stat Ass. 2001;96:746–74.
21. Shtarkov YM. Universal sequential coding of single messages. Probl Inf Transm (Translated from). 1987;23(3):3–17.
22. Rissanen J, Tabus I. Kolmogorov's structure function in mdl theory and lossy data compression. In: Adv. Min. Descrip. Length Theory Appl. Cambridge, MA: MIT Press; 2005. p. 10.
23. Kontkanen P, Myllymäki P. A linear-time algorithm for computing the multinomial stochastic complexity. Inf Process Lett. 2007;103(6):227–33.
24. Roos T, Silander T, Kontkanen P, Myllymäki P. Bayesian network structure learning using factorized nml universal models. In: Proc. 2008 Information Theory and Applications Workshop (ITA-2008). IEEE Press; 2008.
25. Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P. Causal inference using graphical models with the r package pcalg. J Stat Soft. 2012;47(11):1–26.
26. Kalisch M, Bühlmann P. Robustification of the pc-algorithm for directed acyclic graphs. J Comput Graph Stat. 2008;17(4):773–89.
27. Scutari M. Learning Bayesian Networks with the bnlearn R Package. J Stat Soft. 2010;35(3):1–22.
28. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinforma. 2006;7(Suppl 1):7.
29. Meyer PE, Lafitte F, Bontempi G. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinforma. 2008;9:461.
30. Tsamardinos I, Brown LE, Aliferis CF. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Mach Learn. 2006;65(1):31–78.
31. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005;308(5721):523.
32. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol. 2007;3:78.
33. Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, Bansal M, et al. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. Cell. 2009;137(1):172–81.
34. Oram SH, Thoms JAI, Pridans C, Janes ME, Kinston SJ, Anand S, et al. A previously unrecognized promoter of lmo2 forms part of a transcriptional regulatory circuit mediating lmo2 expression in a subset of t-acute lymphoblastic leukaemia patients. Oncogene. 2010;29:5796–5808.
35. Cleveland S, Smith S, Tripathi R, Mathias E, Goodings C, Elliott N, et al. Lmo2 induces hematopoietic stem cell like features in t-cell progenitor cells prior to leukemia. Stem Cells. 2013;31(4):882–94.
36. Moignard V, Macaulay I, Swiers G, Buettner F, Schütte J, Calero-Nieto F, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nat Cell Biol. 2013;15:363–72.

37. Spooner CJ, Cheng JX, Pujadas E, Laslo P, Singh H. A recurrent network involving the transcription factors pu.1 and gfi1 orchestrates innate and adaptive immune cell fates. Immunity. 2009;31(4):576–86.
38. Zohren F, Souroullas G, Luo M, Gerdemann U, Imperato M, Wilson N, et al. The transcription factor lyl-1 regulates lymphoid specification and the maintenance of early t lineage progenitors. Nat Immunol. 2012;13(8):761–9.
39. Li L, Jothi R, Cui K, Lee J, Cohen T, M. Gorivodsky IT, et al. Nuclear adaptor ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. Nat Immunol. 2011;12:129–136.
40. Chan WYI, Follows GA, Lacaud G, Pimanda JE, Landry JR, Kinston S, et al. The paralogous hematopoietic regulators lyl1 and scl are coregulated by ets and gata factors, but lyl1 cannot rescue the early scl–/– phenotype. Blood. 2006;109(5):1908–1916.
41. Chowdhury AH, Ramroop JR, Upadhyay G, Sengupta A, Andrzejczyk A, Saleque S. Differential transcriptional regulation of meis1 by gfi1b and its co-factors lsd1 and corest. PLoS ONE. 2013;8(1):53666. doi:10.1371/journal.pone.0053666.
42. Göttgens B, Nastos A, Kinston S, Piltz S, Delabesse ECM, Stanley M, et al. Establishing the transcriptional programme for blood: the scl stem cell enhancer is regulated by a multiprotein complex containing ets and gata factors. The EMBO J. 2002;21(12):3039–050. doi:10.1093/emboj/cdf286.
43. Dash D, Druzdzel MJ. A hybrid anytime algorithm for the construction of causal models from sparse data. In: Proceedings of the Fifteenth International Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann; 1999. p. 142–9.
44. Cano A, Gomez-Olmedo M, Moral S. A score based ranking of the edges for the pc algorithm. In: Proceedings of the European Workshop on Probabilistic Graphical Models (PGM); 2008. p. 41–8.
45. Claassen T, Heskes T. A bayesian approach to constraint based causal inference. In: In Proc. of the 28th Conference on Uncertainty in Artificial Intelligence (UAI). Burlington, MA: Morgan Kaufmann; 2012. p. 207–16.
46. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6:461–4.
47. Bouckaert RR. Probabilistic network construction using the minimum description length principle. In: Symbolic and Quantitative Approaches to Reasoning and Uncertainty (Clarke M, Kruse R, Moral S, eds). Berlin, Germany: Springer; 1993. p. 41–8.
48. Szpankowski W. Average Case Analysis of Algorithms on Sequences. New York, NY: John Wiley & Sons; 2001.
49. Kontkanen P, Buntine W, Myllymäki P, Rissanen J, Tirri H. Efficient computation of stochastic complexity In: C. Bishop, B. Frey, editors. Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics; 2003. p. 233–8.
50. Kontkanen P. Computationally efficient methods for mdl-optimal density estimation and data clustering. 2009. PhD thesis. Helsinki University Print. Finland.

DEUXIÈME PARTIE

# Apprentissage de modèles graphiques avec variables latentes

CHAPITRE 5

# GRAPHES ANCESTRAUX

## 5.1 Introduction aux graphes ancestraux

Comme nous l'avons introduit dans la section 1.1, il est possible de reconstruire un CPDAG à partir des indépendances conditionnelles, lorsque le système observé satisfait les hypothèses de fidélité et de suffisance causale. Dans les cas pratiques en revanche, la suffisance causale est rarement satisfaite impliquant que le CPDAG reconstruit ne représente pas correctement la structure de génération des données sous-jacente, à moins bien sûr que la possibilité de l'existence de causes communes cachées ne soit explicitement considérée. Il faudrait alors explicitement introduire toutes les variables latentes possibles sans aucune contrainte sur la topologie du réseau ou sur le nombre de causes communes cachées, ce qui crée un espace de recherche infini ([46]). Afin de contourner ce problème, la classe des graphes ancestraux a été introduite par Richardson et Spirtes en 2002 ([34]). Nous allons voir que cette classe de modèles graphiques préserve les relations dites **ancestrales** contenues dans le véritable DAG sous-jacent, malgré la présence de variables latentes ou de sélection.

## 5.2 Complications dues aux variables latentes et de sélection

Il existe deux types de variables cachées pouvant aboutir à l'inférence d'associations erronées lorsque l'on reconstruit un modèle causal : les causes communes (variables **latentes**) et les variables dites de **sélection**. Les variables latentes correspondent à des variables non mesurées, influençant directement plusieurs variables contenues dans les données d'observation. Les variables de sélection, en revanche, induisent un biais en restreignant les échantillons observées à une sous population précise, non représentative de la population générale. Étant donnée que le biais induit par ces variables correspond à une

problématique de design expérimental, on considère que toute indépendance ou association doit toujours être envisagée comme étant conditionnée sur les variables de sélection.

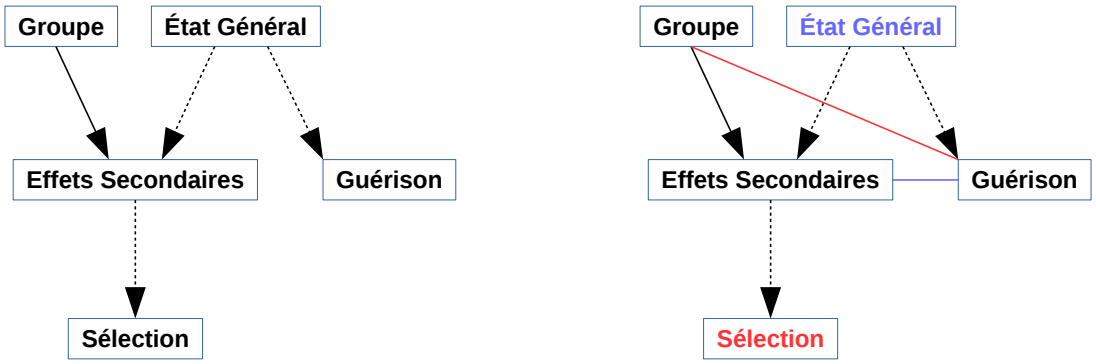
### 5.2.1 Corrélations erronées : Illustration

L'exemple suivant ([18], traduit de l'anglais) permet d'illustrer les effets potentiels de ces deux types de variables sur la reconstruction d'un modèle graphique (fig.5.1).
ESSAI RANDOMISÉ D'UN MÉDICAMENT INEFFICACE POUVANT PROVOQUER DES EFFETS SECONDAIRES IMPORTANTS :
*"Le graphe représente un essai clinique randomisé ayant pour sujet un traitement inefficace entraînant de désagréables effets secondaires. On inclut aléatoirement chaque patient soit dans le groupe contrôle, soit dans le groupe recevant le traitement. Les sujets auxquels le médicament est administré souffrent d'effets secondaires désagréables, avec une sévérité variable en fonction de leur état général (les patients les plus malades souffrant le plus des effets secondaires). Les patients pour lesquels la sévérité des effets secondaires est suffisante sont ceux qui vont le plus probablement abandonner l'étude. Le fait de rester ou non dans l'étude devient alors la variable de sélection. Puisque seuls les patients qui sont les plus malades et qui prennent le médicament sont les plus susceptibles de quitter l'étude, les patients appartenant au groupe auquel le traitement est administré ont tendance à être en meilleure santé que les patients du groupe contrôle. Pour finir, l'état général détermine la rapidité à laquelle le patient guérit."*

Sur la figure 5.1, la fausse corrélation, résultant du biais biais de sélection induit par le fait que seuls les patient **sains** à qui le traitement est administré ne quittent pas l'étude, est représentée en rouge. On note également la corrélation erronée (en bleue), due à la variable latente $EtatGénéral$ cause commune des variables $Effets\ secondaires$ et $Guérison$.

### 5.2.2 Effets causaux erronés

Une autre problématique concerne la découverte d'effets causaux erronés. La figure **??** représente un DAG comprenant 3 variables observées $\{X_i, X_j, X_k\}$ et deux variables latentes $\{L_a, L_b\}$. Parmi les variables observées, la seule indépendance est $X_i \perp\!\!\!\perp X_j$ ce qui amène à penser que le DAG sous-jacent est la v-structure $X_i \rightarrow X_k \leftarrow X_j$. Ce CPDAG suggère que $X_i$ et $X_j$ sont deux causes de $X_k$ alors qu'aucun chemin dirigé n'existe entre ces variables. Comme indiqué dans la section 5.1, les graphes ancestraux sont dans ce cas




A – Mécanisme Causal

B – Corrélations erronées

FIGURE 5.1 – **Corrélations erronées dues à une cause commune latente et à une variable de sélection** – Les variables $\{Groupe,\ Effets\ secondaires,\ Guérison\}$ sont les variables observées, tandis que $Etat\ Général$ est une cause commune latente et que $Sélection$ est une variable de sélection. (**B**) – le lien rouge souligne la fausse corrélation due à la variable de sélection et le lien bleu figure la corrélation erronée due à la cause commune $Etat\ Général$.

plus appropriés pour représenter les interactions entre les variables observées avec la représentation $X_i \leftrightarrow X_k \leftrightarrow X_j$. Cette représentation est appelée graphe ancestral maximal (MAG) et représente le DAG sous-jacent de la figure 5.2 en tenant compte des variables latentes. Par conséquent, la pointe de flèche doit être lue comme "pas une cause de", ce qui signifie que ni $X_i$ ni $X_j$ ne sont des causes de $X_k$ ou d'une variable de sélection. De même pour $X_k$ qui n'est donc ni la cause de $X_j$, ni celle de $X_i$.

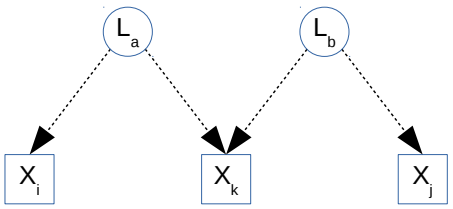



FIGURE 5.2 – **Exemple de DAG avec des variables latentes** – $\{X_i, X_j, X_k\}$ sont les variables observées, tandis que $\{L_a, L_b\}$ sont des variables latentes

## 5.3 Graphes ancestraux : Notations et terminologie

Cette section détaille les notations spécifiques à la classe des graphes mixtes à laquelle appartiennent les graphes ancestraux, introduits précédemment. Un graphe mixte comprend deux types de terminaisons de liens, la tête de flèche ($>$) et la queue de flèche ($-$). Comme nous l'avons déjà vu précédemment, trois types de liens sont alors possibles : **dirigé** ($\rightarrow$), **non dirigé** ($-$) et **bi-dirigé** ($\leftrightarrow$). Le symbole $*$ est utilisé pour représenter le fait que n'importe quelle terminaison est possible pour le lien donné, nous n'y ferons cependant que rarement référence dans le présent manuscrit.
Dans un graphe mixte, deux nœuds $X_i$ et $X_j$ reliés par n'importe quel type de lien sont adjacents. Si $X_i \rightarrow X_j$, $X_i$ est un **parent** de $X_j$ ; dans le cas où $X_i \leftrightarrow X_j$, $X_i$ et $X_j$ sont **partenaires** (littéralement époux en anglais) ; enfin $X_i$ et $X_j$ sont **voisins** dans le cas où $X_i - X_j$. Une chaîne de nœuds **adjacents et ditincts** (sont éventuellement deux extrémités) est un **chemin**. Un chemin de $X_0$ vers $X_p$ est dit **dirigé** si $\forall 0 \leq i < p$, $X_i$ est un parent de $X_{i+1}$. Si $X_p$ est également parent de $X_0$, on retrouve la notion abordée précédemment de **cycle dirigé**. Dans le cas où ce cycle contient un lien bi-dirigé, c'est un **cycle presque dirigé** (*almost directed cycle*).
La variable $X_i$ est considéré comme un **ancêtre** de $X_j$ (noté $an_{\mathcal{G}}(X_j)$) s'il existe un chemin dirigé de $X_i$ vers $X_j$, qui est alors appelé **descendant** de $X_i$ (noté $de_{\mathcal{G}}(X_j)$). Si on suppose que les données d'observations ont été générées par un DAG et que le système observé n'est pas complet (il existe des variables latentes et/ou de sélection), il est possible de représenter l'ensemble des indépendances conditionnelles ainsi que les interactions causales héritées de ce DAG sous-jacent grâce au **graphe ancestral maximal** (MAG) reconstruit sur les variables observées ([46]). Pour être ancestral, un graphe mixte $\mathcal{G}$ doit respecter les conditions suivantes :

- $\mathcal{G}$ ne comprend aucun cycle dirigé.
- $\mathcal{G}$ ne comprend aucun cycle presque dirigé.
- Pour tout lien non-dirigé $X_i - X_j$, $X_i$ et $X_j$ n'ont ni parents, ni partenaires.

La première de ces conditions est bien entendu liée au fait que l'on a supposé que le graphe sous-jacent est un DAG et donc acyclique. La seconde condition, couplée à la première, permet d'insister sur l'interprétation différente de la tête de flèche dans les graphes ancestraux, c'est-à-dire que celle-ci représente une "non-cause". Par exemple, dans un graphe ancestral $X_i \rightarrow X_j$, la tête de flèche en $X_j$ indique que $X_j$ n'est pas la cause de $X_i$. En revanche, l'absence de cette tête de flèche du côté de $X_i$ apporte

une information de causalité en indiquant que $X_i$ est la cause de $X_j$, ou d'une variable intermédiaire cachée. Par conséquent la queue de flèche est moins informative que la tête de flèche dans le cas d'un biais de sélection. Enfin la troisième condition permet une représentation correcte de l'effet de sélection en simplifiant également l'adaptation des graphes ancestraux. Pour résumer, le graphe ancestral permet de préserver parmi les variables observées les interactions ancestral contenues dans le DAG sous-jacent.

## 5.4 Equivalence de Markov, définition du PAG

De la même manière que pour les DAG, on peut définir l'équivalence de Markov pour deux MAGs de la façon suivante :

### 5.4.1 Définition de l'équivalence de Markov de deux MAGs

Deux MAGs $\mathcal{M}_{\mathcal{G}_1},\ \mathcal{M}_{\mathcal{G}_2}$ sont équivalents markoviens si $\forall$ ensemble de nœuds disjoints $X, Y, Z \subseteq \mathbf{O}$, les ensembles $X$ et $Y$ sont indépendants conditionnés sur $Z$ dans $\mathcal{M}_{\mathcal{G}_1}$ et dans $\mathcal{M}_{\mathcal{G}_2}$. Autrement dit, l'ensemble de séparation d'une paire donnée est identique dans les deux MAGs considérés.
Par conséquent, tous les membres d'une classe d'équivalence de Markov d'un MAG donné ont les mêmes adjacences ainsi que les mêmes terminaisons **invariables** (tête ($>$) ou queue ($-$) de flèche). En effet, dans une classe d'équivalence de MAGs donnée, une ou les deux terminaisons de certains liens peuvent varier d'un MAG à l'autre. La marque $\circ$ sert à signaler ces terminaisons variables.
Les terminaisons invariables et les liens reconstruits à partir des indépendances conditionnelles peuvent être représentés au sein d'un **Graphe Ancestral Partiel** (PAG). Cette représentation autorise 6 sortes de liens différents : $-$, $\rightarrow$, $\leftrightarrow$, $-\circ$, $\circ\!-\!\circ$ et $\circ\!\rightarrow$. Un lien bi-dirigé souligne la présence d'une cause commune cachée, tandis qu'un lien non-dirigé reflète la présence d'une variable de sélection.

### 5.4.2 Définition du PAG

Formellement, le PAG peut être définit de la façon suivante :
On note $\mathcal{G}$ le DAG pour lequel $\mathbf{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$, avec $\mathbf{O}$ l'ensemble des variables observées, $\mathbf{L}$ l'ensemble des variables latentes et $\mathbf{S}$ l'ensemble des variables de sélection. Ces trois

ensembles sont disjoints. Prenons $\mathcal{M}$ un graphe mixte reconstruit sur l'ensemble de variables observées **O**. $\mathcal{M}$ est un PAG représentant $\mathcal{G}$ si et seulement si les quatre conditions suivantes sont respectées :

1. si $X_i$ et $X_j$ ne sont pas adjacents dans $\mathcal{M}$, alors il existe un ensemble $Y \subseteq \mathbf{O}$ $\{X_i, X_j\}$ tel que $X_i \perp\!\!\!\perp X_j | \mathbf{Y} \cup \mathbf{S}$
2. si $X_i$ et $X_j$ sont reliés par un lien non dirigé dans $\mathcal{M}$, alors il n'existe pas $Y \subseteq \mathbf{O}$ $\{X_i, X_j\}$ tel que $X_i \perp\!\!\!\perp X_j | \mathbf{Y} \cup \mathbf{S}$. Autrement dit, $X_i$ et $X_j$ sont dépendants quelque soit le set de conditionnement issu de l'ensemble des variables observées.
3. Si $X_i \circ\!\!\rightarrow X_j$ dans $\mathcal{M}$ alors $X_j \notin an_{\mathcal{M}}(X_i \cup S)$. Cette condition illustre l'interprétation différente de la tête de flèche dans les graphes ancestraux : $X_j$ n'est pas la cause de $X_i$.
4. Si $X_i \circ\!\!-\!\!\circ X_j$ dans $\mathcal{M}$ alors $Xj \in an_{\mathcal{M}}(X_i \cup S)$. Ce qui veut dire que l'absence de pointe de flèche implique que $X_j$ est la cause, soit de $X_i$, soit d'une variable de sélection cachée.

Pour illustrer cette définition, on peut se référer à la figure 5.2. Le modèle graphique $X_i \circ\!\!\rightarrow X_k \leftarrow\!\!\circ X_j$ est le PAG reconstruit sur l'ensemble **O** des variables observées représentant la class d'équivalence markovienne des MAGs du DAG de cette figure 5.2. Dans cette représentation, les têtes de flèche encadrant $X_k$ signifient que $X_k$ ne peut être ni la cause de $X_i$, ni celle de $X_j$, ni bien sûr d'une variable de sélection intermédiaire. Le $\circ$ représente quant à lui l'incertitude concernant la possibilité pour $X_i$ et $X_j$ d'être des causes de $X_k$ ou d'une variable de sélection intermédiaire.

## 5.5 Découverte d'adjacences dans un PAG

Comme introduit dans la section 5.2, les variables cachées empêchent l'inférence de certaines indépendances conditionnelles parmi les variables observées. L'exemple de la figure 5.5, tiré de [13], illustre une dépendance causale erronée due à la présence d'une variable latente. Ici, la variable latente $L$ empêche la découverte de l'indépendance entre $X_i \perp\!\!\!\perp X_j | X_k$ dans l'ensemble des variables observées **O** empêchant ainsi la suppression du lien entre $X_i$ et $X_j$, bien que $X_i$ n'ait pas d'influence sur $X_j$ dans le DAG $\mathcal{G}$ (on rappelle que pour $\mathcal{G}, V = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$). Cependant, étant donnée la définition du PAG $\mathcal{M}$ (voir 5.4.2), une adjacence entre deux variables observées peut être décidée en prenant en compte les variables cachées dans le cas où :

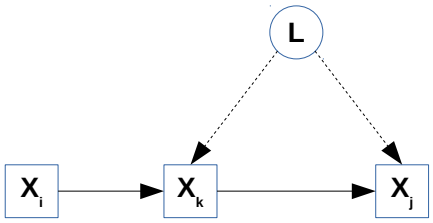



FIGURE 5.3 – **Exemple d'un DAG $\mathcal{G}$ sur $\mathbf{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$.** $\{X_i, X_k, X_j\} \subseteq \mathbf{O}$ avec **O** correspondant à l'ensemble des variables observées, $\{L\} \subseteq \mathbf{L}$ avec **L** correspondant à l'ensemble des variables latentes et $\emptyset \subseteq \mathbf{S}$ correspondant aux variables de sélection (absentes ici).

$$\left.\begin{array}{r} X_i \notin an_{\mathcal{M}}(X_j \cup \mathbf{S}) \\ \text{and} \\ X_i \perp\!\!\!\perp X_j | Y \cup \mathbf{S} \end{array}\right\} \Rightarrow X_i \perp\!\!\!\perp | Y' \cup \mathbf{S}$$

avec

- $Y \subseteq \mathbf{O}\ \{X_i, X_j\}$
- $Y' \subset$ D-SEP$\{X_i, X_j\}$ ou $Y' \subset$ D-SEP$\{X_j, X_i\}$

Par conséquent, pour inférer une adjacence entre $X_i$ et $X_j$ dans le PAG $\mathcal{M}$, il est suffisant de trouver un ensemble de séparation $Y'$ à partir d'un ensemble D-SEP de $X_i, X_j$ donné, plutôt que d'essayer tous les sous-ensembles $Y \subseteq \mathbf{O} \backslash \{X_i, X_j\}$. Cependant, les ensembles D-SEP dont la définition est étroitement liée à la notion de **chemins induits**, ne peuvent être déduits à partir des indépendances conditionnelles observées ([13], [42]). Un "super-ensemble" appelé D-SEP Possible et contenant l'ensemble D-SEP a été proposé par Spirtes et al., 2000 ([41]). Ce D-SEP possible se définit de la façon suivante :

### 5.5.1 Concept de D-SEP possible

Prenons le DAG $\mathcal{G}$ avec $\mathbf{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$. Notons $\mathcal{C}$ le PAG représentant $\mathcal{G}$. Un nœud $X_k$ appartient au D-SEP Possible$(X_i, X_j)$, noté $pds_{\mathcal{C}}(X_i, X_j)$ si et seulement s'il existe un chemin $\pi$ entre $X_i$ et $X_j$ dans $\mathcal{C}$ tel que pour tous les sous-chemins $\langle X_m, X_l, X_h \rangle$ de $\pi$, $X_l$ est la pointe d'une v-structure sur ce sous-chemin dans $\mathcal{C}$ ou $\langle X_m, X_l, X_h \rangle$ est un

triangle dans $\mathcal{C}$.

Comme nous le détaillerons plus loin, cette notion d'ensemble de variables D-SEP Possible peut permettre de reconstruire le PAG représentant de la classe d'équivalence de Markov des DAGs sous-jacent, lorsque l'on suppose que les hypothèse de fidélité et de suffisance causale ne sont pas respectées, c'est-à-dire que l'on souhaite tenir compte de l'effet de variables latentes et/ou de sélection lors de la reconstruction du modèle.

CHAPITRE 6

# INFÉRENCE DE PAGS : ALGORITHME FCI (FAST CAUSAL INFERENCE)

## 6.1 Contexte et Principe

Comme nous l'avons introduit dans la section 5.2, la présence de variables cachées, telles que les variables latentes ou de sélection, entraîne des complications lorsque l'on cherche à reconstruire un modèle graphique causal à partir des variables observées. L'algorithme FCI (Fast Causal Inference) ([42], [41]) permet la reconstruction du représentant de la classe d'équivalence de Markov à partir d'indépendances conditionnelles déduites des variables observées, tout en tenant compte de la présence de variables latentes ou de sélection. L'algorithme 2 détaille le fonctionnement de FCI. La première étape consiste à appliquer l'algorithme `PC` sur un graphe complet (entièrement connecté) sur l'ensemble de variables observées **O**. Ce premier squelette est ensuite partiellement orienté : seules les v-structures sont déduites des ensembles de séparation renvoyé par `PC`. C'est au cours de la troisième étape que les corrélations dues aux variables latentes sont explorées de façon complémentaire, en ne cherchant les ensembles de séparations que dans le D-SEP Possible (voir section 5.5.1) (Algorithme 2, étape 3). Les orientations du squelette définitif obtenu à l'issue de cette étape sont ensuite initialisée à $\circ\!\!-\!\!\circ$. L'orientation finale est finalement effectuée, en deux étape : la règle $R_0$ identifie les v-structures, qui sont ensuite propagées en suivant les règles $R_1$ à $R_{10}$ ([46]).

La figure 6.1, adaptée de l'article dont le preprint est présenté en 7.2, illustre la reconstruction sur les variables observées du PAG représentant la classe d'équivalence de Markov du DAG (fig.6.1A). Les indépendances causales ont uniquement été déduites des variables observées, en utilisant le principe du D-SEP possible.

En pratique, le lien entre les variables $Z$ et $T$ n'a pas été supprimé après l'application de l'algorithme PC (fig. 6.1B) car la variable $W$, requise pour inférer cette indépendance

conditionnelle ne fait pas partie des adjacences des variables impliquées. Cependant, l'application du principe du D-SEP possible durant l'étape 3 de l'algorithme permet de rechercher des nœuds plus distant pour trouver l'ensemble de séparation. Étant donné le fait que $pdsc_{\mathcal{C}_2}(T,Z)\backslash\{T,Z\} = pdsc_{\mathcal{C}_2}(Z,T)\backslash\{T,Z\} = \{X,Y,W\}$, l'algorithme FCI retrouve l'indépendance conditionnelle : $Z \perp\!\!\!\perp T|\{X,Y,W\}$, permettant d'inférer correctement le PAG.

Comme nous le verrons, notamment sur les tests comparatifs réalisés sur différents réseaux benchmarks, la recherche des D-SEP possible peut très rapidement devenir gourmande sur le plan computationnel. En effet, ces D-SEP possibles sont susceptibles de devenir très importants, même si on considère des réseaux relativement peu connectés (voir figure 1F, article Verny et al., *Plos Comp. Biol.* 7.2).

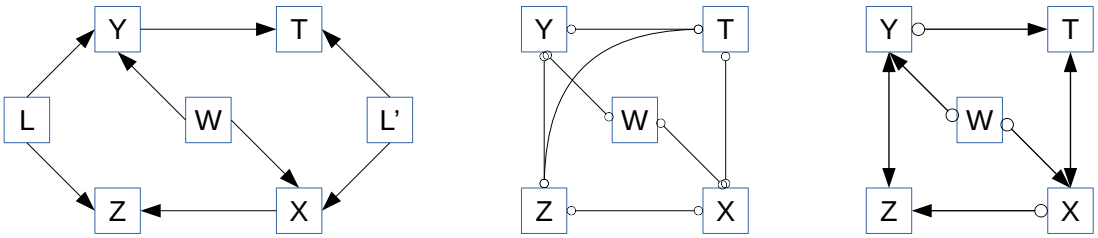



FIGURE 6.1 – **Inférence d'un PAG $\mathcal{C}$ par FCI** – Ici, $\mathbf{O} = \{T,X,Y,Z\}$, $\mathbf{L} = \{L,L'\}$ et on suppose $\mathbf{S} = \emptyset$.

## 6.2 Amélioration : Really Fast Causal Inference (RFCI)

L'une des dernières améliorations en date de l'algorithme FCI, **RFCI** ([6]) tente d'accélérer la reconstruction du modèle par rapport à FCI, qui souffre comme on l'a dit d'une complexité exponentielle due à la taille des D-SEP possibles.

L'approche RFCI suit les mêmes étapes que l'algorithme FCI, mais en introduisant des modifications pour les étapes 2 et 3. En pratique, durant l'étape 2, l'algorithme effectue des tests supplémentaires pour supprimer certains liens superflus avant d'orienter les

v-structures (**Règle des triplets ouverts** ([6])) en s'affranchissant de la règle obligeant les ensembles de séparation à faire partie des adjacences des variables d'un nœud. La seconde modification, appelée **règle des chemins discriminants**, consiste à l'utilisation de tests supplémentaires de découvertes d'indépendances conditionnelles pour les triangles correspondant aux conditions d'application de la règle $R_4$ d'orientation. L'application de ces tests permets là encore de supprimer des liens superflus.
La suppression des liens grâce à l'application de ces deux règles permet de :

1. Eviter la recherche des D-SEP possibles sur les liens correspondants
2. Diminuer la taille de ces D-SEP possibles en rendant le graphe moins connecté

En pratique cependant, les études réalisées sur les différents réseaux benchmarks n'ont montré que peu de différence entre le temps d'exécution de FCI et celui de RFCI ; montrant que les modifications introduites ne sont que peu utilisées, car probablement rarement applicables.

Nous verrons dans le chapitre suivant que, contrairement aux algorithmes FCI et RFCI, `MIIC` n'est que peu concerné par l'impact de la recherche de variables cachées sur son temps d'exécution.

**Algorithme 2 :** Agorithme FCI

**Entrée :** Données d'observation des variables **O** ; un classement $ordre(\mathbf{O})$ des variables ; un seuil de confiance $\alpha$

**Sortie :** (un graphe mixte interprétable comme) un PAG $\mathcal{C}$ ([45])

**0. Initialisation**

Débuter avec un graph complet non dirigé $\mathcal{C}$ avec tous les liens de la forme $\circ\!\!-\!\!\circ$

**1. Inférence du squelette initial, ensembles de séparation. Sortie :**$\mathcal{C}_1$

Appliquer l'algorithme PC pour trouver le squelette initial $\mathcal{C}_1$ ainsi que les ensembles de séparation

**2. Inférence des orientations initiales. Sortie :**$\mathcal{C}_2$

**pour tous** *les triplets ouverts de* $\mathcal{C}_1$ **faire**

| $R_0 : \{X_i *\!\!-\!\!\circ X_k \circ\!\!-\!\!* X_j \ \& \nexists\, X_i - X_j \ \& \ X_k \notin Sep_{X_iX_j}\} \Rightarrow \{X_i *\!\!\rightarrow X_k \leftarrow\!\!* X_j\}$

**fin**

**3. Squelette final, ensembles de séparation. Sortie :**$\mathcal{C}_3$

**pour** *tous les sommets* $X_i \in \mathcal{C}_2$ **faire**

| Calculer $pdsc_{\mathcal{C}_2}(X_i, \bullet)$
| **pour** *tous les sommets* $X_j \in adj_{\mathcal{C}_2}(X)$ **faire**
| | initialiser $l = 0$
| | **tant que** $X_iX_j \in adj_{\mathcal{C}_2}$ *avec* $|pdsc_{\mathcal{C}_2}(X_i, \bullet)\backslash\{X_j\}| \geq l$ **faire**
| | | **tant que** $\exists\, Y \subset pdsc_{\mathcal{C}_2}(X_i, \bullet)\backslash\{X_j\}$ *non considéré, avec* $|Y| = l$ **faire**
| | | | **si** $X_i \perp\!\!\!\perp X_j | \{Y, S\}$ *au niveau de conf.* $\alpha$ **alors**
| | | | | **le lien** $X_iX_j$ n'est pas essentiel et est **supprimé**
| | | | | **ensemble de séparation** de $X_iX_j$ : $\mathrm{Sep}_{X_iX_j} = \{Y\}$
| | | **fin**
| | **fin**
| | $l = l + 1$
| **fin**

**fin**

Initialiser tous les liens de $\mathcal{C}_3$ à $\circ\!\!-\!\!\circ$

**4. Orientation finale. Sortie :** $\mathcal{C}_4$

**pour tous** *les triplets ouverts de* $\mathcal{C}_3$ **faire**

| $R_0 : \{X_i *\!\!-\!\!\circ X_k \circ\!\!-\!\!* X_j \ \& \nexists\, X_i - X_j \ \& \ X_k \notin Sep_{X_iX_j}\} \Rightarrow \{X_i *\!\!\rightarrow X_k \leftarrow\!\!* X_j\}$

**fin**

**5. Propagation. Sortie :** $\mathcal{C}_5$

**tant que** $\exists$ *une orientation pouvant être propagée* **faire**

| Appliquer les règles $R_1$ à $R_{10}$ tirées de ([46])

**fin**

CHAPITRE 7

# MULTIVARIATE INFORMATION-BASED INDUCTIVE CAUSATION (MIIC)

## 7.1 MIIC : Avancées

### 7.1.1 Détection des causes communes cachées (variables latentes)

3off2, la version précédente de notre méthode de reconstruction de réseaux publiée dans *BMC Bioinformatics*[2], ne tient pas compte de la différence entre les interactions dues aux variables latentes et les interactions directes. MIIC offre cette possibilité, grâce à l'élargissement de la recherche des contributeurs à toutes les variables du réseau considéré, pas seulement les voisins. Ceci permet de traiter correctement les reconstructions de réseaux complexes comprenant des variables latentes, tels que celui présenté en fig. 1C du preprint *Plos Comp Biol* présenté dans la section suivante. La figure suivante présente les performances de MIIC sur le modèle graphique présenté sur la figure 6.1. Les données d'observation ont été simulées en utilisant le logiciel Tetrad (accessible à l'adresse http://www.phil.cmu.edu/tetrad/index.html). On remarque immédiatement sur ces résultats que FCI et MIIC ont de meilleures performances sur les orientations que PC et 3off2, respectivement ; ceci est logique puisque FCI et MIIC autorisent les liens bi-dirigés. On note également les meilleures résultats de FCI et MIIC sur les squelettes, pour les grands jeu de données ($N \geq 750$). En effet, pour ces jeux de données, les algorithmes 3off2 et PC sont incapables de supprimer le lien dû aux corrélations erronées, imputables à la présence des variables latentes.

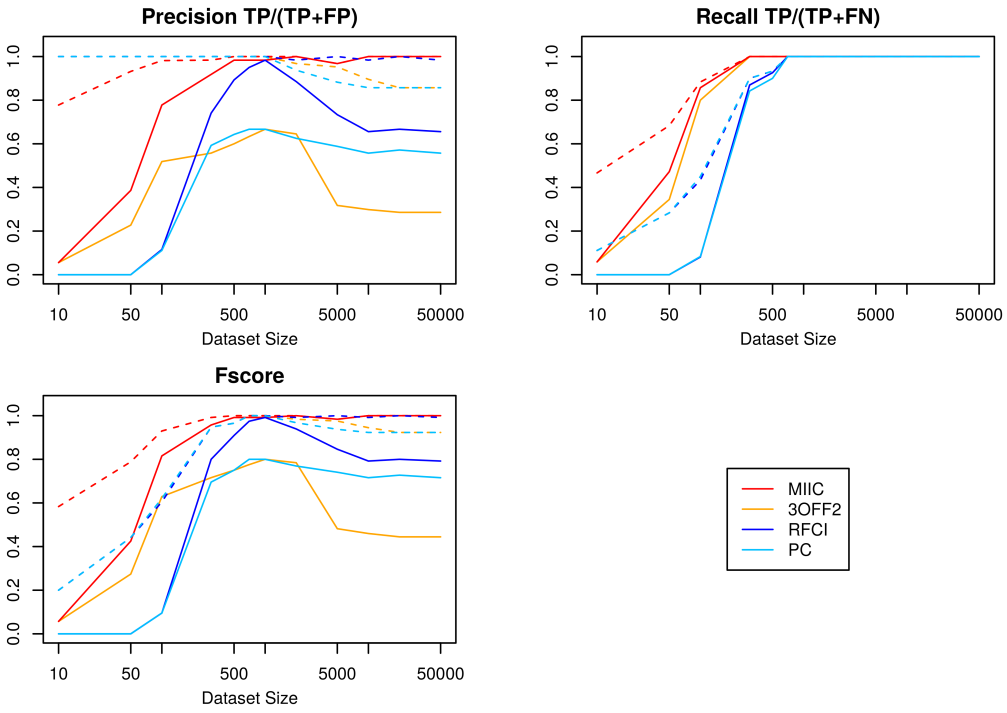



FIGURE 7.1 – **Performances de `MIIC`, `3off2`, FCI, PC sur le PAG présenté sur la figure 6.1** Les lignes en pointillés représentent les performances sur le squelette, tandis que les lignes pleines incluent les orientations.

### 7.1.2 Évaluation quantitative de la "confiance" dans les liens inférés

Le second apport de `miic` par rapport à `3off2` est la possibilité de calculer un indice de confiance spécifique à chaque lien inféré et d'utiliser cette quantité pour effectuer un filtrage. Ce ratio se présente sous la forme suivante pour un lien entre deux variable $X$ et $Y$ :

$$C_{XY} = \frac{P_{XY}}{\langle P_{XY}^{rand} \rangle} \tag{7.1}$$

où $P_{XY}$ représente la probabilité de ne pas conserver le lien en utilisant les données originale, et $\langle P_{XY}^{rand} \rangle$ représente la moyenne des probabilité de ne pas conserver ce lien calculées sur les données randomisées plusieurs fois. Plus la quantité $C_{XY}$ est proche de $0$, plus la confiance dans le lien $XY$ est donc importante (faible probabilité de ne pas conserver le lien avec les données originales *vs* forte probabilité de le supprimer lorsqu'on randomise les données). Les études de l'impact de ce seuil de confiance sur les réseaux benchmarks montrent (fig. 7.2), comme attendu, une augmentation de la précision au détriment du

recall, avec un impact relativement faible sur le Fscore (moyenne harmonique de la précision et du recall).

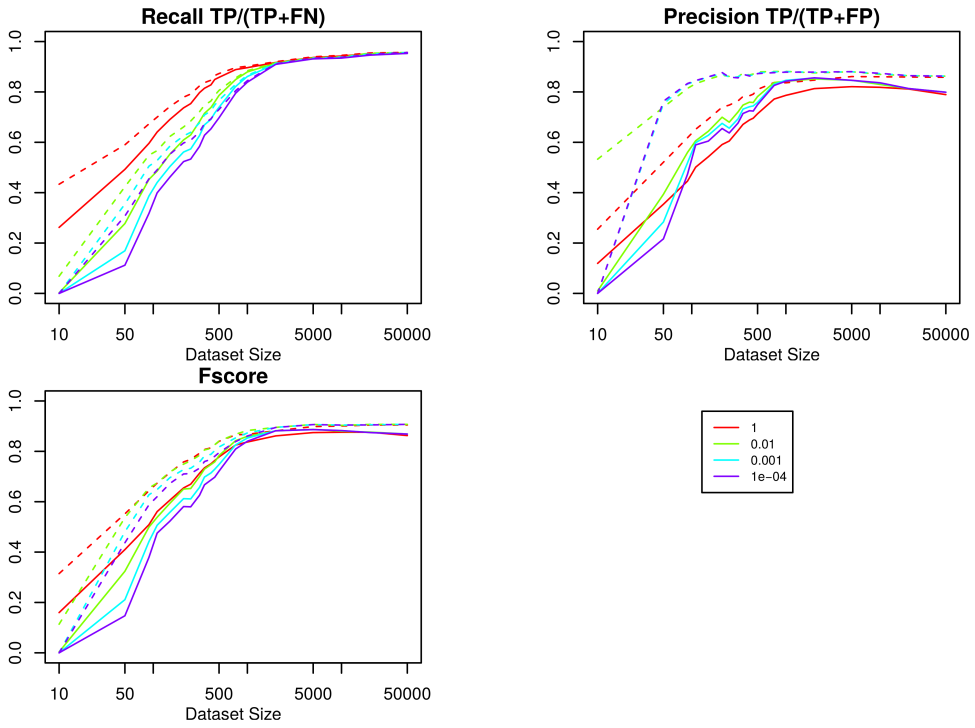



FIGURE 7.2 – **Variations des performances de** MIIC **à différents seuils de confiance** – De la même manière que précédemment, les lignes pointillées représentent les résultats sur le squelette, tandis que les lignes pleines représentent les résultats sur le graphe dirigé.

### 7.1.3 Nombre d'échantillons efficaces

Pour finir, la reconstruction d'un modèle graphique par MIIC suppose des échantillons indépendants les uns des autres, contrairement à des jeux de données *time-series* par exemple.

Pour corriger ce type de biais dans les jeux de données analysés, nous avons analysé les fonctions **d'autocorrélation** prenant en compte les corrélations observées dans des données de séries temporelles. Il est en effet possible d'analyser l'autocorrélation dans ce type de données en raison du fait qu'elle est prévisible dans le sens probabiliste, puisque les valeurs observées dans le future dépendent du passé. Dans le cas d'une décroissance exponentielle de cette autocorrélation au sein du jeu de données, il est possible d'estimer la correction nécessaire grâce à **l'autocorrélation de premier ordre**, qui indique la

similarité entre deux observations successives. Elle est définie par :

$$r_1 = \frac{\sum_{t=1}^{N-1}(x_t - \bar{x})(x_{t+1} - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2} \tag{7.2}$$

où $N$ représente le nombre d'échantillons du jeu de données, $x_t$ le vecteur des valeurs des variables pour l'échantillon $t$, $x_{t+1}$ le vecteur des valeurs des variables pour l'échantillon suivant et $\bar{x}$ le vecteur des valeurs moyennes des variables. Ce coefficient peut ensuite être utilisé pour déterminer le nombre d'échantillons indépendants du jeu de donnée grâce à la relation :

$$N^*_{eff} = N\frac{1 - r_1}{1 + r_1} \tag{7.3}$$

Avec $\frac{1-r_1}{1+r_1}$ le facteur de correction qui évolue tel que présenté sur la figure 7.1.3, en fonction de $r_1$.

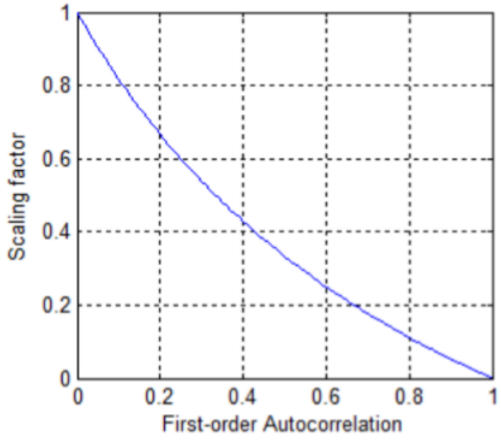



FIGURE 7.3 – **Évolution du facteur correctif en fonction du coefficient d'autocorrélation de premier ordre** – Pour un $r_1 = 0.1$, on aura donc un facteur de correction $\simeq 0.8$, ce qui signifie que seuls $80\%$ des échantillons sont réellement indépendants

Dans le but d'obtenir une reconstruction viable dans le cas d'échantillons corrélés, `MIIC` est donc capable de détecter une autocorrélation à décroissance exponentielle et d'appliquer la correction définie ci-dessus dans ce cas.

### 7.1.4 Calcul des signes des interactions

Le calcul du signe d'une interaction est un aspect qui nous a semblé indispensable pour faciliter l'analyse de réseaux reconstruits sur des données expérimentales réelles. En effet, il n'est pas rare que lorsque ces signes ne sont pas calculés, des confusions soient faites dans l'interprétation de la causalité inférée. En partant de cette idée, nous avons implémenté dans MIIC une façon relativement simple de calculer ce signe, basée sur le **coefficient de corrélation partielle**, défini de la manière suivante pour un lien $X_iX_j$ avec un contributeur $A_i$ :

$$r_{X_iX_j|A_i} = \frac{r_{X_iX_j} - r_{X_iA_i} \times r_{X_jA_i}}{\sqrt{1 - r^2_{X_iA_i}} \times \sqrt{1 - r^2_{X_jA_i}}} \tag{7.4}$$

où $r_{X_iX_j}$ est le coefficient de corrélation entre les variables $X_i$ et $X_j$, $r_{X_iA_i}$ est le coefficient de corrélation entre la variable $X_i$ et le contributeur $A_i$ et $r_{X_jA_i}$ est le coefficient de corrélation entre la variable $X_j$ et le contributeur $A_i$. Le signe de ce coefficient permet de déterminer si l'association entre les deux variables est négative (représentée en bleue sur les réseaux de ce manuscrit) ou positive (en gris).

## 7.2 Publication de MIIC dans *Plos Computational Biology*

L'algorithme MIIC ainsi que les différents points présentés ci-dessus ont fait l'objet d'un article, accepté par la revue *Plos Computational Biology*. Le preprint de cet article est présenté dans cette section.

# Learning causal networks with latent variables from multivariate information in genomic data

Louis Verny[1,2☯], Nadir Sella[1,2☯], Séverine Affeldt[1,2☯¤a], Param Priya Singh[1,2¤b], Hervé Isambert[1,2*]

**1** Institut Curie, PSL Research University, CNRS, UMR168, 26 rue d'Ulm, 75005 Paris, France
**2** Sorbonne Universités, UPMC Univ Paris 06, 4, Place Jussieu, 75005 Paris, France

☯These authors contributed equally to this work.
¤a Current address: LIPADE, University of Paris Descartes, 45 rue des Saints Pères, 75006 Paris, France
¤b Current address: Department of Genetics, Stanford University, Palo Alto, USA
* herve.isambert@curie.fr

## Abstract

Learning causal networks from large-scale genomic data remains challenging in absence of time series or controlled perturbation experiments. We report an information-theoretic method which learns a large class of causal or non-causal graphical models from purely observational data, while including the effects of unobserved latent variables, commonly found in many genomic datasets. Starting from a complete graph, the method iteratively removes dispensable edges, by uncovering significant information contributions from indirect paths, and assesses edge-specific confidences from randomization of available data. The remaining edges are then oriented based on the signature of causality in observational data. The approach and associated algorithm, `miic`, outperform earlier methods on a broad range of benchmark networks. Causal network reconstructions are presented at different biological size and time scales, from gene regulation in single cells to whole genome duplication in tumor development as well as long term evolution of vertebrates.

## Author summary

The reconstruction of causal networks from genomic data is an important but challenging problem. Predicting key regulatory interactions or genomic alterations at the origin of human diseases can guide experimental investigation and ultimately inspire innovative therapy. However, causal relationships are difficult to establish without the possibility to directly perturb the organisms' genome for ethical or practical reasons. Besides, unmeasured (latent) variables may be hidden in many genomic datasets and lead to spurious causal relationships between observed variables. We propose in this paper an efficient computational approach, `miic`, that overcomes these limitations and learns causal networks from non-perturbative (observational) data in the presence of latent variables. In addition, we assess the confidence of each predicted interaction and demonstrate the enhanced robustness and accuracy of `miic` compared to alternative existing methods. This approach can be applied on a wide range of datasets and provide new biological insights on regulatory networks from single cell expression data or genomic alterations during tumor development. `Miic` is implemented in an R package freely available to the scientific community under a General Public License.

# Introduction

Network reconstruction methods have become ubiquitous to analyze large-scale information-rich data from the latest genomic technologies. Recently, methodological advances in the field have been seeking to learn causal relationships using time series or controlled perturbation experiments [1, 2]. However, such strategies can be technically impracticable or costly, if not unethical, in many biological contexts.

Alternatively, graphical models can be learned by simply observing enough random variations in unperturbed data, as for the reconstruction of gene regulatory networks from single-cell gene expression data. However, most methods based on this principle, such as Bayesian search-and-score [3], sparse inverse covariance estimation [4], maximum entropy [5] or diffusion map [6] methods, assume as underlying models either causal networks with only directed edges or non-causal networks with only undirected edges. Thus, they cannot uncover nor rule out causality in observational data. By contrast, constraint-based methods [7–10], which identify structural constraints corresponding to all dispensable edges in a graph, can in principle uncover causality from purely observational data. Advanced constraint-based methods [9, 10] reconstruct Markov equivalent models of a broad class of "ancestral graphs" [11], that include undirected ($-$), directed ($\rightarrow$) and possibly bidirected ($\longleftrightarrow$) edges originating from latent common causes, $L$, unobserved in the available data (*i.e.* $\leftarrow\!\!-L\!\!-\!\!\rightarrow$). However, constraint-based methods are often not robust on small datasets and have algorithmic complexity issues when including unobserved latent variables [9–12]. Yet, latent variables are commonly found in many real applications, as in the case of an unobserved transcription factor $TF$ co-regulating two co-expressed genes, *i.e.* $G_1 \leftarrow\!\!- TF \!\!-\!\!\rightarrow G_2$ (see example of single cell transcriptomics in Fig. 2). These unobserved variables should not be ignored in practice, as they actually impact the causal relationships between observed variables, leading to spurious causal association between co-regulated genes $G_1$ and $G_2$ in the previous example. While the algorithmic difficulties of constraint-based methods have so far limited their applicability in practice, understanding cause-effect relationships [13] remains of primary interest to model complex biological systems and anticipate their response to environmental changes or genetic alterations.

We report here an information-theoretic method, that simultaneously circumvents the complexity and robustness issues of constraint-based approaches, and demonstrate its applicability to real biological data. The method builds on an earlier information-theoretic approach [14], in order to *i)* include latent variables, a notorious conceptual and algorithmic difficulty in causal network reconstruction [9–13], and *ii)* provide an edge specific confidence assessment of retained edges, which lacks in traditional constraint-based methods. Both aspects are important in practice to reconstruct robust networks from actual biological data. The approach is applied to reconstruct causal networks from a variety of genomic datasets at different biological size and time scales, from single cells to organisms and entire phyla.

# Results

### Background: Signature of causality and unobserved latent variables in observational data

Our information-theoretic method for network reconstruction is based on the analysis of multivariate information [14–19], $I(X;Y;Z;\cdots)$, which extends the concept of mutual information [20] beyond two variables, $I(X;Y) = \sum_{x,y} p(x,y)\log(p(x,y)/p(x)p(y))$, where $p(x)$, $p(y)$ and $p(x,y)$ are the measured probability distributions of single or joint

variables $X$ and $Y$ from the available data $\mathcal{D}$ (see Materials and Methods). Most importantly, unlike two-point mutual information, $I(X;Y)$, which cannot distinguish causal from non-causal relations between variables $X$ and $Y$, multivariate information involving more than two points, $I(X;Y;Z;\cdots)$, may imply cause-effect relationships between the underlying variables, S1 File.

In fact, the signature of causality in purely observational data is associated to a unique correlation pattern involving at least three variables [13, 21]: it concerns two mutually (or conditionally) independent variables, $I(X;Y) = 0$, which are therefore not connected to each other, yet both connected to a third variable $Z$, Fig. 1A. This situation entails the orientations of a 'v-structure' or 'unshielded' collider, $X \to Z \leftarrow Y$, because the edges $XZ$ and $YZ$ cannot be undirected, nor $Z$ be a cause of $X$ or $Y$, as these alternative graphical models imply correlations that would contradict independence between $X$ and $Y$. V-structures are the hallmark of causality in observational data: networks with v-structures are necessary causal, while causal models without v-structures can be shown to be equivalent to their undirected counterparts from the viewpoint of observational data.

Beyond v-structures, colliders may also be found in series along a collider path, as in $X \to Z \longleftrightarrow Y \leftarrow W$, Fig. 1B & C, where the bidirected edge, $Z \longleftrightarrow Y$, indicates that $Z$ is not a cause of $Y$ nor $Y$ a cause of $Z$. It implies that the correlation between $Z$ and $Y$ is due to at least one latent common cause, $L$, unobserved in the available dataset, $Z \dashleftarrow L \dashrightarrow Y$, as outlined above. Hence, statistical dependencies and independencies in purely observational data can, in principle, provide structural constraints for network reconstruction as well as information on causal relationships across observed and possibly unobserved latent variables. These results underline the wealth of information which cannot be captured from two-point correlations only.

## An information-theoretic method to learn causal networks with latent variables

The signature of causality and unobserved latent variables in multi-point correlation statistics enables to rephrase constraint-based methods [7–10] within an information-theoretic framework. Constraint-based approaches, sketched in Fig. 1D, start from a fully connected network and proceed by iteratively removing dispensable edges between variables $X$ and $Y$ for which a conditional independence can be found, *i.e.* $I(X;Y|\{A_i\}) = 0$ (Fig. 1D, step 1). This rationale of constraint-based methods can be interpreted from an information perspective [22], using the generic decomposition of mutual information, $I(X;Y)$, relative to the set of variables $\{A_i\}$,

$$I(X;Y) = I(X;Y;\{A_i\}) + I(X;Y|\{A_i\}), \tag{1}$$

where $I(X;Y;\{A_i\})$ can be seen as the global indirect contribution of $\{A_i\}$ to $I(X;Y)$ and $I(X;Y|\{A_i\})$ as the remaining (direct) contribution (see Eq. 8 in Materials and Methods). Conditional independence, $I(X;Y|\{A_i\}) = 0$, then implies that $\{A_i\}$ is a 'separation set' which intercepts all indirect paths contributing to the total mutual information, *i.e.* $I(X;Y) = I(X;Y;\{A_i\})$. In practice, however, conditional mutual information cannot be exactly zero for finite datasets but the probability that the $XY$ edge should be removed can be estimated from the available data as, $P_{XY} \sim \exp(-N\, I(X;Y|\{A_i\}))$, up to some normalization constant, where $N$ is the number of independent samples (S1 File). The undirected network 'skeleton', resulting from the removal of all dispensable edges, is then partially directed by orienting all v-structures (Fig. 1D, step 2), based on the signature of causality, outlined above, and propagating these orientations on downstream edges (Fig. 1D, step 3), based on specific

propagation rules consistent with ancestral graphs [23].

The main computational complexity of constraint-based methods is to uncover a valid combination of contributing nodes $\{A_i\}$ for each dispensable edge $XY$. In absence of latent variables, the combinatorial search can be restricted to the sole neighbors of $X$ or $Y$, which are sufficient to intercept all information contributions from indirect paths [7,8]. However, this efficient algorithm cannot be used in the presence of latent variables, as collider paths may require to extend the combinatorial search for conditioning set $\{A_i\}$ to non-adjacent variables of $X$ and $Y$ [9], as illustrated in Fig. 1C. In practice, this intrinsic difficulty stemming from latent variables has been addressed through more complex algorithmic approaches, such as the `FCI` algorithm [9] and its more recent approximate variant, `RFCI` [10]. Beyond algorithmic complexity issues, traditional constraint-based methods are also known to be highly sensitive to sampling noise inherent to finite datasets and are not robust on typical datasets of interest (*e.g.* expression data of 30 to 40 genes measured in a few hundreds to thousands of single cells [24], see application and Fig. 2 below).

The present algorithmic approach, `miic` (multivariate information-based inductive causation), circumvents the complexity and robustness issues of standard constraint-based methods by avoiding to directly tackle the combinatorial search of complete separation sets. Instead, it progressively collects, one-by-one, their most likely contributors, $\{A_i\}_n = \{A_1, A_2, \cdots, A_n\}$, based on a quantitative score for each pair of variables $XY$ (S1 File). The global indirect contribution is then obtained iteratively as,

$$I(X;Y;\{A_i\}_n) = I(X;Y;\{A_i\}_{n-1}) + I(X;Y;A_n|\{A_i\}_{n-1}), \quad (2)$$

where $I(X;Y;A_n|\{A_i\}_{n-1}) > 0$, corresponds to the contribution of the most likely $n$th variable $A_n$ after collecting the first $n-1$ most likely contributors, $\{A_i\}_{n-1}$ (see Eq. 10 in Materials and Methods). We demonstrate in the current study that this iterative framework, which proved to be robust to sampling noise in absence of latent variables [19], can in fact be extended to include latent variables by collecting the contributors $\{A_i\}$ within the whole set of observed variables, instead of amongst the sole neighbors of $X$ and $Y$ in absence of latent variables [14]. This simple approach to include latent variables circumvents the algorithmic complexity of standard constraint-based methods [9,10], while improving ten to hundred folds their performance in both prediction accuracy and running time, as discussed in the next section.

### Algorithmic performance on causal and non-causal benchmark datasets

We have assessed the performance of `miic` on a broad range of causal and non-causal benchmark networks from real-life as well as simulated datasets from $P \simeq 30$ up to 500 variables and $N = 10$ up to $50,000$ independent samples (Materials and Methods). The causal benchmark networks, which include an increasing fraction (0% to 20%) of hidden latent variables, are derived using partially observed Bayesian networks, that is, considering some variables as hidden. These unobserved variables are usually present in many real applications and cannot be ignored in practice, as they actually impact the causal relationships between observed variables, as illustrated in Fig. 1B-D. The non-causal benchmark datasets have been obtained from Monte Carlo sampling of Ising-like interacting networks sharing approximately the same two-point direct correlations with real-life benchmark causal networks but lacking causality. Monte Carlo sampling leads, however, to significant correlations between successive samples, which needs to be taken into account through an effective number of independent samples (Materials and Methods).

Reconstructed causal networks have been compared to *partial ancestral graphs*

(PAGs) [23], which are the representatives of the Markov equivalent class of all ancestral graphs consistent with the conditional independences in the available data. In practice, benchmark PAGs have been derived by hidding some variables in benchmark directed acyclic graphs (DAG) using the `dag2pag` function of the `pcalg` package with slight modifications [25, 26]. The alternative inference methods used for comparison with `miic` are the `FCI` algorithm [9] and its recent approximate variant `RFCI` [10] implemented in the `pcalg` package [25, 26]. The results obtained with `FCI` and `RFCI` are in fact very similar and we only present here comparisons with the more recent `RFCI` algorithm [10]. `RFCI`'s results are shown for an adjustable significance level $\alpha = 0.01$ and using the *stable* implementation of the skeleton learning algorithm, as well as the *majority rule* for the orientation and propagation steps [27], which give overall the best results. The results have been evaluated in terms of running time, as well as, Precision (or positive predictive value), Recall or Sensitivity (true positive rate), and F-score, which is the harmonic mean of Precision and Recall (Materials and Methods). Precision, Recall and F-score have been derived for the undirected skeleton of the networks (dashed lines in Fig. 1E) or taking into account edge orientations (solid lines in Fig. 1E).

The results on benchmark networks are presented in Fig. 1E, Fig. 1F, as well as S1 Fig, S2 Fig, S3 Fig, S4 Fig, S5 Fig, S6 Fig and S7 Fig. `Miic` outperforms classical constraint-based approaches, including its advanced approximate variant `RFCI`, Fig. 1E, especially on networks with many underlying parameters. It achieves significantly better or comparable results with much fewer samples (Fig. 1E, S1 Fig, S2 Fig and S3 Fig), and is typically ten to hundred times faster (Fig. 1F). In addition, `miic`'s ability to learn complex ancestral networks, which require conditioning on non-adjacent variables, can be directly demonstrated on the example of Fig. 1C network, S4 Fig. The complexity of `miic` algorithm, while difficult to evaluate exactly, proves to be linear in terms of sample size (Fig. 1F) and quadratic in terms of network size for sparse graphs irrespective of the inclusion of latent variables (S5 Fig). By contrast, traditional constraint-based methods exhibit roughly quadratic complexity in terms of sample size (Fig. 1F) and much steeper complexity scaling in terms of network size, especially when latent variables are included [12]. Furthermore, no causality is predicted by `miic` for non causal datasets, even from small effective numbers of independent samples (Materials and Methods and S6 Fig and S7 Fig). This underlines `miic` accuracy to uncover true causality.

### Edge confidence assessments

This information-theoretic method and its algorithmic implementation (S1 Software) are very general and can be applied to a wide range of datasets, provided a sufficient number of independent samples is available. We report here the results obtained with genomic datasets spanning a broad range of biological size and time scales from single cells and tissues to organisms and entire phyla. In addition to including latent causal variables, we have also assessed the confidence of predicted edges with an edge specific confidence ratio $C_{XY} = P_{XY}/\langle P_{XY}^{\text{rand}}\rangle$, where $P_{XY}$ is the probability to remove the $XY$ edge, introduced above, and $\langle P_{XY}^{\text{rand}}\rangle$ the average of the same probability after randomizing the datasets for each variable (see Materials and Methods, and S1 File section 2.2 for details). Hence, the lower $C_{XY}$, the higher the confidence on the $XY$ edge, which can be used to retain only high confidence edges in the predicted networks.

Interestingly, the effect of confidence filtering on the reconstruction of benchmark networks (S8 Fig & S9 Fig) demonstrates that the filtering of individual edges improves the Precision of the reconstruction (at the expense of its Sensitivity or Recall) not only for the network skeleton, as expected, but also for the network orientations, while retaining overall similar F-scores. This demonstrates the interest and consistency of

using such confidence filtering to obtain an enhanced and tunable precision of the reconstructed networks for real biological applications. Indeed, an enhanced precision might be desirable in many practical applications for which the correctness of predicted edges is more important than the occasional dismissal of less certain edges. All network reconstructions presented in Fig. 2, Fig. 3 & Fig. 4 have been obtained with an edge specific confidence $C_{XY} < 10^{-3}$, while network skeletons obtained before edge filtering are displayed in S11 Fig, S14 Fig and S15 Fig.

The general three-step reconstruction scheme of `miic` (*i.e.* Step 1- graph skeleton, Step 2- edge filtering, Step 3- edge orientation) is also sensitive to the fine tuning of other algorithmic parameters such as the complexity criterion introduced to estimate finite size effects. All results presented in this paper have been obtained with the decomposable Normalized Maximum Likelihood (NML) criterion introduced in [28, 29], which was shown to yield significantly better results than more traditional BIC/MDL criterion on benchmark networks, especially on small datasets, leading to simultaneous improvements in both recall and precision [19]. Choosing the BIC/MDL instead of NML criterion in the three genetic network applications, Fig. 2, Fig. 3 & Fig. 4, leads to somewhat sparser reconstituted networks including 82% to 100% of initial edges, yet no additional edges (*i.e.* consistent with a lower recall), and 66% to 75% conserved edge orientations (*i.e.* identical —, →, ← and ↔ edges), see S1 Table.

### Analysis of expression data in single cells

At cellular level, we reconstructed regulatory networks from single cell expression data at the time of endothelial and hematopoietic differentiations from the primitive streak cells of the mouse early embryo, Fig. 2A. This concerns the formation of primitive erythroid cells, a distinct and transient red blood cell lineage arising directly from mesodermal progenitors with restricted hematopoietic potential [30], by contrast to the highly studied definitive erythroid cells which arise from multipotent hematopoietic stem cells.

The dataset for this application is from Moignard *et al* [24] and includes the expression of 33 transcription factors (TFs) along with 13 non-TF genes (markers) in 3,934 single cells extracted at 4 different times of the mouse embryo development (days E7.0, E7.5, E7.75 and E8.25), Fig. 2A-C and S10 Fig. The cells extracted from E8.25 were also divided by the authors in two different pools: potential endothelial precursors and potential hematopoietic precursors based on the expression of the *Runx1* hematopoietic marker. Gene expression was collected using single cell qRT-PCR and binarized by the authors, leading to two-state (on / off) expression levels in the available dataset. Pooling all cells together regardless of their developmental timing (from day E7.0 to E8.25), we first analyzed their population heterogeneity using principal component analysis (PCA), Fig. 2B, and K-means clustering, Fig. 2C. Three main cell populations are identified and can be interpreted, based on gene functional classification (Materials and Methods), as progenitor, endothelial precursor and hematopoietic precursor populations, whose relative proportions vary from E7.0 to E8.25, Fig. 2C.

The network predicted by `miic`, Fig. 2D, includes 75 edges with $C_{XY} < 10^{-3}$ out of 82 edges in the unfiltered skeleton, S11 Fig. The differentiation bifurcation between endothelial and hematopoietic precursors, seen through principal component (Fig. 2B) and clustering (Fig. 2C) analyses, also clearly appears in the reconstructed regulatory network, Fig. 2D, after labelling hematopoietic specific TFs (in red), endothelial TFs (in purple) and common TFs expressed in both precursor lineages (in blue), Materials and Methods. In fact, most predicted regulatory interactions across lineage specific TFs correspond to regulatory inhibitions (in blue), which might originate either from direct regulatory repressions or possibly through indirect 'ancestor' regulations involving unobserved intermediary TFs. In addition, a number of known regulatory interactions

are correctly predicted in the inferred network, Fig. 2D, such as *Ikaros* → *Gfi1b* and *Ikaros* → *Lyl1* [31], *Tal1* → *Fli1* and *Tal1* → *Lmo2* [32] as well as *HoxB4* → *Erg* (with opposite orientation) and *Sox7* → *Erg* [24]. Yet, there are also many predicted regulations in `miic` network that have not been reported so far as well as a number of regulations documented in definitive erythroid cells [32] that appear to be missing in primitive erythroid cells (*e.g. Est1* → *Tal1*, *Sfpi1* → *Tal1* and *Sfpi1* → *Myb*). These results suggest a number of testable predictions, including five bidirected edges consistent with the absence of direct regulations reported between these genes. Indeed, bidirected edges imply the necessity to invoke unobserved latent co-regulators between such genes. In particular, the unmeasured *Gata2* expression is possibly implicated in the co-regulation of *Erg* ↔ *Lyl1*, based on an earlier study [33]. Hence, beyond the consistency with earlier reports as well as testable predictions, `miic` results may also help pinpoint possible latent regulators unobserved in Moignard *et al*'s study [24], such as regulators specific to the initial progenitor cells, not yet committed to either hematopoietic or endothelial lineages and accounting for about 70% of analyzed cells at day E7.0, Fig. 2C.

### Analysis of genomic and ploidy alterations in breast tumors

At tissue and organismal levels, we analyzed genomic alterations on breast tumors from the online Catalog of Somatic Mutations in Cancer (COSMIC) datase [34], Fig. 3A-C.

The dataset, which contains 807 samples without predisposing *BRCA1/2* germline mutations, includes somatic mutations (from whole exome sequencing) and expression level information for 91 genes. These 91 genes have been selected based on earlier studies on mutation and/or expression alterations in breast cancer, Materials and Methods. Gene non-synonymous mutation status is binarized (yes / no) and gene expression status is categorized as under-, normal- or over-expressed by the COSMIC database. S12 Fig provides the distribution of altered expressions and S13 Fig the distribution of mutations for the 91 genes of interest. In addition to gene mutations and altered expression levels, we also integrated information on sample average ploidy, provided by the COSMIC database (release v76) and discretized the clearly bimodal ploidy distribution (Fig. 3B) with ploidy $< 2.7$ considered as diploid cells and ploidy $\geqslant 2.7$ taken as tetraploid cells, in agreement with COSMIC convention [34]. Among the 807 samples, 401 correspond to diploid tumoral cells and 398 to tetraploid tumoral cells (8 samples have no ploidy information). As expected, *TP53, RB1* and *PTEN* tumor suppressors tend to be mutated, downregulated or lost, especially in tetraploid tumors, Fig. 3B & C, which also exhibit significant normalized expression alterations, Fig. 3C.

The network predicted by `miic` is shown Fig. 3D. We first note that, due to the limited numbers of samples (N=807) and recurrent gene mutants (Fig. 3C and S13 Fig), most gene mutations are not confidently linked to any altered expression levels (compare Fig. 3D with edge confidence $C_{XY} < 10^{-3}$ to the unfiltered skeleton, S14 Fig), with the notable exceptions of *TP53* and *RB1* mutations, which have a significant impact on gene expressions, Fig. 3D. Interestingly, the overall effect of tetraploidization on normalized gene expression, Fig. 3C, is predicted to be largely indirect and mediated by *TP53* mutations which lead to dysregulation of mitosis controling genes, such as the under-expression of *PPP2R2A* [35] and over-expression of *AURKA* and *CENPA* genes. In addition, tetraploidy and *TP53* mutations tend also to be concomitant with over-expression of metabolic (*GMPS*) and cell-growth modulating genes (*TSPYL5*, *NDRG1* and *FOXM1*) [36], favoring tumor progression and metastasis, as well as higher expression of *APOBEC3B*, which promotes mutational heterogeneity within tumors and,

thereby, their drug resistance through subclonal selection [37]. Hence, `miic` results provide a direct link between the long-known incidence of *TP53* mutations in (breast) cancer and the tetraploidization of tumor cells. These results, supported by a number of recent reports [35, 37–40], shed light on the poor prognosis associated with tetraploid tumors and their resistance to chemotherapy [40]. This presumably occurs as tetraploid cells can exploit their genome redundancy and heterogeneity to evolve resistance strategies under drug treatments, Fig. 3A.

Interestingly, this dynamics of tetraploid tumors in the course of cancer progression and treatment echoes the success of tetraploid species in the course of eukaryote evolution. Indeed, genome doubling events, possibly associated to environmental changes, have repeatedly led to successful evolutionary radiations of biodiverse subphyla, such as the vertebrates and the flowering plants [41], although the underlying selection mechanism has remained a matter of debate [41–44].

### Analysis of two rounds of tetraploidization in vertebrate evolution

We have investigated with `miic` this long term evolution following the two rounds of tetraploidization that occurred in early vertebrates some 500 million years ago, Fig. 4A. While long lost species and subphyla cannot be directly studied, the genetic make up of extant vertebrates provides an information-rich data on the selection processes at work since these ancient genome duplications. In particular, we aimed at identifying the genomic properties potentially responsible for the biaised retention of 'ohnolog' gene duplicates [45] retained from these genome duplications in early vertebrates.

We obtained 20,415 protein-coding genes in the human genome from Ensembl (v70) and collected information on the retention of duplicates originating either from the two whole genome duplications at the onset of vertebrates ('ohnolog') or from subsequent small scale duplications ('SSD') as well as copy number variants ('CNV'), Fig. 4B and S1 Data [45]. 5,504 ohnolog genes retained from the two rounds of whole genome duplications (WGDs) in the common vertebrate ancestor were obtained from the 'Ohnologs' server based on multi-species comparison of synteny [45]. All the small scale duplicates (SSDs) in the human genome were obtained from Ensembl Compara using BioMart [46], and were restricted to the 4,506 genes duplicated after the WGDs. Genes with copy number variants (CNVs) were obtained from the Database of Genomic Variants [47]. A total of 5,185 genes were identified to be CNV genes as their entire coding sequence fell within one of the CNV regions in this database.

We then collected information on the genomic properties of these 20,415 human genes, including their sequence conservation ('Ka/Ks ratio'), protein autoinhibitory folds and participation to protein complexes, their expression levels across tissues, association with dominant or recessive diseases and susceptibility to cancer mutations as well as their essentiality for development and reproduction, see Materials and Methods.

The resulting causal network, predicted by `miic`, relates the origin of duplicated genes in the human genome (*i.e.* 'ohnolog', SSD or CNV gene duplicates) to their genomic properties and association to diseases, Fig. 4C. The reconstructed network implies that the retention of ohnolog duplicates is more directly linked to their susceptibility to dominant mutations and protein autoinhibitory folds than other genomic properties such as dosage balance constraints in protein complexes [42], gene essentiality or expression levels, which do not exhibit direct links to ohnolog retention, Fig. 4C, even on the network skeleton obtained before edge confidence filtering, S15 Fig. Hence, `miic` analysis based on observational data provides an independent confirmation as well as significant extension of earlier reports based on correlations between two or

three genomic properties [43] and on simple population genetic models [48]. All together, these results support an evolutionary retention of ohnologs by purifying selection through dominant diseases in tetraploid species (consistent with the retention of ohnologs with low Ka/Ks ratio, Fig. 4C, indicating sequence conservation) while small scale duplicated genes have been retained through positive selection (consistent with their higher Ka/Ks ratio, Fig. 4C, indicative of underlying adaptation).

# Discussion

We report in this paper a novel information-theoretic method that learns a broad class of network models including latent causal effects from purely observational data, that is, in absence of time series or controlled intervention experiments, which can be technically impractical, costly or unethical to obtain in many biological contexts.

The methodology of our approach is quite general and follows a three-step scheme:

- Step 1- Find a graph skeleton taking into account latent variables.
- Step 2- Remove weakly supported edges based on a confidence criterion.
- Step 3- Determine edge orientations based on the signature of causality.

While resembling traditional constraint-based methods such as `FCI`, `miic` is in fact designed to be much faster and more robust to finite sample size through greedy algorithmic strategies based on quantitative information-theoretic scores at each algorithmic step, *i.e.* Step 1: iterative collection of most likely contributors based on an contributor ranking scheme, Step 2: filtering of weakly supported edges through an edge-specific confidence assessment, and Step 3: successive orientation of the remaining edges based on decreasing orientation probabilities.

Unlike earlier robust methods for network reconstruction [3–6], this general scheme circumvents the need to choose between causal and non-causal graphical models *a priori*, as the most appropriate class of models is directly learned from the available data. In addition, the approach can uncover the effect of unobserved latent variables, a notorious conceptual and algorithmic difficulty in causal network reconstruction [13]. Yet, latent variables are usually present in many real applications and cannot be ignored in practice, as they actually impact the causal relationships between observed variables.

More specifically, `miic` relies on the analysis of multivariate information [14–19], which extends the concept of mutual information to more than two variables. In practice, `miic` integration of constraint-based methods within an information-theoretic framework leads to greatly improved performances in both prediction accuracy (Fig. 1E) and running time (Fig. 1F) as well as favorable scalings in terms of sample size (Fig. 1F) and network size (S5 Fig). The likelihood ratio formalism also enables to derive an edge specific confidence index, $C_{XY}$, which allows to filter predicted edges to obtain an enhanced and tunable precision of the reconstructed networks. This might be desirable in many applications for which the correctness of predicted edges is more important than the occasional dismissal of less certain edges.

We have used `miic` to reconstruct causal networks from a variety of genomic datasets at different biological size and time scales, from gene regulation in single cells (Fig. 2) to whole genome duplication in tumor development (Fig. 3) as well as long term evolution of vertebrates (Fig. 4). In all these applications, `miic` provides testable predictions and new biological insights summarized below:

*i)* on the hematopoietic / endothelial differentiation network (Fig. 2), `miic` results shed lights on the regulatory interactions in primitive erythropoietic differentiation for which much less is known compared with definitive erythropoiesis [30].

We predict, in particular, the central role of regulators such as *Ikaros* in the hematopoietic precursor population, and *Sox7* and *Erg* in the endothelial precursor population, as well as the causal effects of unobserved latent variables such as the transcription factor *Gata2*;

*ii)* on the development of breast cancer, `miic` network reconstruction (Fig. 3) highlights the direct association between tetraploidization and *TP53* mutations, by contrast with earlier studies on non-cancerous cell lines [40, 49] but in agreement with findings on actual tumors and their resistance to treatments [38, 40]. These results are also consistent with the high incidence of tetraploid tumors in patients with *BRCA1/2* germline mutations [50];

*iii)* finally, concerning the impact of whole genome duplications in vertebrate evolution, `miic` results (Fig. 4) refute the general view in the field on the retention of ohnologs through dosage balance constraints [42]. Instead, `miic` multivariate analysis demonstrates the role of dominant deleterious effects on the retention of ohnologs, which significantly extends and confirms earlier reports based on correlations between two or three genomic properties [43, 44] and independent population genetic results based on first-principles evolutionary models [48].

Beyond the three genomic network reconstructions presented in this paper (Fig. 2, Fig. 3 and Fig. 4), we anticipate that this information-theoretic approach may help uncover cause-effect relationships in other information-rich datasets from different fields of biological interest, such as developmental biology, neuroscience, clinical data analysis and epidemiology. The causal network learning tool, `miic`, is implemented in an R-package software with open source code and freely available under a General Public License (S1 Software).

# Materials and Methods

## Application

### Gene functional classification in hematopoiesis / epithelial differentiation

The early hematopoiesis single cell transcription data come from Moignard *et al.*, 2015 [24]. The expression of 33 TFs and 13 non-TF genes (markers) have been obtained by single cell qRT-PCR and binarized (on/off) by the authors. The 33 TFs can be classified into 3 categories related to their function, using the Mouse Genome Database [32] as well as the TF expressions at the different time points in the original experiment [24]:

- "Hematopoietic": This group gathers the TFs for which we found a function in hematopoietic differentiation, without finding any evidence of a role in endothelium formation in the litterature. The corresponding genes linked to hematopoietic function are: *Eto2, Sfpi1/PU.1, Runx1, Nfe2, Myb, Mitf, Ikaros, Gfi1b, Gfi1, Gata1.*
- "Endothelial": For these genes, the main function found in the litterature is in endothelial development. The corresponding genes linked to endothelial function are: *Ets2, Erg, Tbx3, Tbx20, Sox7, Sox17, Notch1, HoxB4.*

- "Common": These TFs have been shown to be involved in both hematopoietic and endothelial differentiation. The corresponding genes linked to both hematopoietic and endothelial functions are: *Fli1, Etv6, Etv2, Ets1, Tal1, Meis1, Mecom, Lyl1, Lmo2, Ldb1, Hhex.*

### Signature gene set in breast cancer progression

The choice of specific genes for monitoring genomic alterations has been guided by earlier studies and breast cancer-specific molecular tests [51], which demonstrate that altered expression profiles can reveal patient overall outcome [52]. In particular, the MammaPrint genomic assay relies on a 70-gene expression profile to assess patient breast cancer recurrence risk [52]. This signature classifies patient either as high-risk or low-risk for long-term development of distant metastasis. The relevance of the MammaPrint 70-gene profile has already been assessed by multiple studies, *e.g.* [52, 53]. Interestingly, although the MammaPrint biomarker genes were selected from a completely data-driven approach, they are enriched with specific cancer hallmarks [54] acquired in the course of tumorigenesis and metastasis progression [55].

In this study, we investigated the interrelations between ploidy, mutation and expression level alterations for 91 genes in breast tumors. Specifically, we first considered the mutation status and expression levels of 50 genes out of the 70 Mammaprint biomarkers for which a hallmark of cancer has been identified [55]. We also considered 18 commonly altered genes in breast cancer (*ERBB2, ESR1, TP53, RB1, MYC, JUN, CDKN2A, BCL2, APOBEC3B, PTEN, MDM2, USP7, UBE3A, SPDYE7P, PLK1, BAX, MET, FOXM1*) [56]. In addition, 23 genes related to ploidy alteration were also included (*TP73, LATS2, MAPK14, CDKN1A, CHEK1, AURKB, AURKA, BRCA1, BRCA2, DUSP5, MST1, PPP1R13L, BIRC3, TGFA, ETS1, ETS2, HIF1A, LDHA, FOXO1, NDRG1, PPP2R1A, PPP2R2A, CCNE1*) [38, 40].

### Genomic properties of ohnolog genes in vertebrates

The genomic properties susceptible to be associated with the retention of 'ohnolog' gene duplicates (as well as SSD and CNV duplicates) in the human genome have been obtained from various resources:

- **Cancer mutations.** Cancer mutation profiles for all the protein coding genes were obtained from the COSMIC database [34]. We counted all the non-synonymous mutations per unit length in all the available samples, and partitioned the 18,538 genes with available mutation information into three equal frequency bins (S1 Data).

- **Disease genes.** Human disease genes were collected from OMIM, GeneCards [57], and from published curated lists [44, 58] and combined to give a total of 7,171 disease genes.

- **Recessive *vs* dominant genes.** Based on the inheritance information from Online Mendelian Inheritance in Man (OMIM) database, we could obtain 981 and 952 genes that were described as autosomal dominant and autosomal recessive genes respectively.

- **Autoinhibition.** Genes with autoinhibitory protein folds were obtained from search and manual curation in PubMed and in various databases (OMIM, SwissProt, NCBI Gene and GeneCards). Additional autoinhibitory candidates

with the domains known to be frequently implicated in autoinhibition (*e.g.* SH3, DH, PH, CH, Drf and Eth domains) were obtained based on the domains identified using HMMER search [59] against Pfam database [60]. This led to a total of 881 genes with autoinhibitory protein folds (S1 Data).

- **Essentiality.** A total of 6,436 1-to-1 mouse orthologs obtained using BioMart and tested for lethality or infertility phenotypes on loss-of-function or knockout mutations in mouse were obtained from the Mouse Genome Informatics database [32]. 2,729 [resp. 3,227] of these 6,436 genes were found to be essential [resp. non-essential] genes in mouse.

- **Protein complex.** A total of 6,119 genes involved in protein complex formation were obtained by combining the protein complexes from Human Protein Reference Database [61], CORUM database [62], the human soluble protein complex census [63], and the human genes belonging to the Gene Ontology term "protein complex" under Cellular Component.

- **Ka/Ks ratio.** We obtained Ka/Ks (or dN/dS) ratios between human and amphioxus (*Branchiostoma floridae*) orthologs using the KaKs_Calculator 2.0 [64]. Ka/Ks ratios were retrieved for a total of 15,508 genes and partitioned into 75% lower ratio $< 0.2$ (*i.e.* more conserved sequences) and 25% higher ratio $\geqslant 0.2$ (*i.e.* rapidly evolving sequences)

- **Expression levels.** Gene expression levels for 78 healthy human tissues and cell types [65] were downloaded from BioGPS [66]. Affimetrix tags were mapped to Ensembl gene IDs using BioMart and annotation provided by BioGPS. Expression levels from different tags for the same gene were averaged after removing the tags that bind to multiple genes. A total of 13,425 genes with an expression level were partitioned into three equal frequency bins based on the their median expression across 78 tissues/cell types.

These genomic properties susceptible to be associated with the retention of 'ohnolog', SSD and CNV gene duplicates are provided as S1 Data.

For each genomic property or combination of properties for which a number of samples presents missing data, multivariate information, such as $I(X;Y|\{A_i\})$, are computed on the number of available samples $N_a$ without missing data for $X$, $Y$ and $\{A_i\}$ variables ($N_a < N$). Finite size corrections are then estimated based on $N_a$ instead of $N$ samples (S1 File). This assumes that the missing data is missing completely at random.

## Methodology

### Ancestral Graphs

The `miic` software reconstructs Markov equivalent models of the broad class of '***ancestral graphs***' [11] which can contain three types of edges, undirected (—), directed ($\rightarrow$) and bidirected ($\longleftrightarrow$) edges, but:

*i)* no directed cycles (*i.e.* $X \rightarrow \rightarrow \cdots \rightarrow \rightarrow Y$ with $X \leftarrow Y$)
*ii)* no almost directed cycles (*i.e.* $X \rightarrow \rightarrow \cdots \rightarrow \rightarrow Y$ with $X \longleftrightarrow Y$)
*iii)* no arrowheads pointing to an undirected edge (*i.e.* $\rightarrow -$ or $\longleftrightarrow -$)

### Multivariate information and most likely information contributors

The `miic` algorithm is an information-theoretic method that learns graphical models by progressively uncovering the information contributions of indirect paths in terms of ***multivariate information***.

The ***multivariate information*** between $p$ variables, $I(X_1;\cdots;X_p)$, is defined through alternating (inclusion-exclusion) sums of multivariate entropies $H(\{X_i\}) = -\sum_{\{x_i\}} p(\{x_i\}) \log p(\{x_i\})$ over all subsets of variables $\{X_i\} \subseteq \{X_1,\cdots,X_p\}$ as [15–17],

$$I(X_1;\cdots;X_p) = \sum_i H(X_i) - \sum_{i<j} H(X_i,X_j) + \sum_{i<j<k} H(X_i,X_j,X_k) - \cdots (-1)^{k-1} \sum_{i_1<\cdots<i_k} H(X_{i_1},\cdots,X_{i_k}) + \cdots (-1)^{p-1} H(X_1,\cdots,X_p) \quad (3)$$

In particular, for $p = 2$ and 3 variables, it yields,

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad (4)$$

$$I(X;Y;A) = H(X) + H(Y) + H(A) - H(X,Y) - H(X,A) - H(Y,A) + H(X,Y,A) \quad (5)$$

where the 3-point information, $I(X;Y;A)$, can be positive or negative unlike the 2-point (mutual) information, $I(X;Y)$, which is always positive [20]. Conditional multivariate information, $I(X_1;\cdots;X_p|A)$, are defined similarly as multivariate information, $I(X_1;\cdots;X_p)$, but in terms of conditional multivariate entropies [18], $H(\{X_i\}|A)$. In particular, conditional mutual information is defined as,

$$\begin{aligned} I(X;Y|A) &= H(X|A) + H(Y|A) - H(X,Y|A) \\ &= -H(A) + H(X,A) + H(Y,A) - H(X,Y,A) \end{aligned} \quad (6)$$

using the definition of conditional entropy [20], $H(X|A) = H(X,A) - H(A)$. Then combining the expressions of $I(X;Y|A)$ and $I(X;Y;A)$ yields a generic decomposition rule relative to a variable $A$ or a set of variables $\{A_i\}_m = \{A_1, A_2, \cdots, A_m\}$ as,

$$I(X;Y) = I(X;Y|A) + I(X;Y;A) \quad (7)$$

$$I(X;Y) = I(X;Y|\{A_i\}_m) + I(X;Y;\{A_i\}_m) \quad (8)$$

and conditioning Eq. 7 on $\{A_i\}_{n-1}$ and setting $A \equiv A_n$ yields,

$$I(X;Y|\{A_i\}_{n-1}) = I(X;Y|\{A_i\}_n) + I(X;Y;A_n|\{A_i\}_{n-1}) \quad (9)$$

which can be combined with Eq. 8, setting $\{A_i\}_m = \{A_i\}_{n-1}$ or $\{A_i\}_n$, to yield the following iterative scheme on the contribution increment of the collected set $\{A_i\}_n$ (see Results),

$$I(X;Y;\{A_i\}_n) = I(X;Y;\{A_i\}_{n-1}) + I(X;Y;A_n|\{A_i\}_{n-1}) \quad (10)$$

As explained in S1 File, only positive information terms, $I(X;Y;A_n|\{A_i\}_{n-1}) > 0$, contribute to the global mutual information between $X$ and $Y$ through the iterative decomposition of Eq. 9,

$$I(X;Y) = I(X;Y;A_1) + I(X;Y;A_2|A_1) + \cdots + I(X;Y;A_n|\{A_i\}_{n-1}) + I(X;Y|\{A_i\}_n) \quad (11)$$

where the most likely contributors $A_n$ after collecting the first $n-1$ contributors $\{A_i\}_{n-1}$ is chosen by maximizing $I(X;Y;A_n|\{A_i\}_{n-1}) > 0$, while taking into account

the finite size $N$ of the dataset (S1 File). The approach provides also a natural ranking of the edges $XY$ of the graph, $R(XY; A_n|\{A_i\}_{n-1})$, based on the likelihood of their best next contributor $A_n$ (Eq. S20 in S1 File).

By contrast, negative information, $I(X;Y;A_n|\{A_i\}_{n-1}) < 0$, do not contribute to $I(X;Y)$ but are the signature of causality in observational data and are used to orient v-structures, such as $X \to A_n \leftarrow Y$ (S1 File).

### Description of `miic` algorithmic pipeline

The implementation of the information-theoretical approach `miic` proceeds in three steps corresponding to the following algorithmic pipeline, Fig. 1D (S1 File):

- Step 1: *Learning skeleton taking into account latent variables*

  Starting from a fully connected undirected graph, `miic` iteratively removes all dispensable edges after collecting one-by-one their most likely contributors $\{A_i\}$ based on the edge ranking order, $R(XY; A_n|\{A_i\}_{n-1})$ (Eq. S20 in S1 File), and using the following pseudocode,

  **Repeat**: take the top edge $XY$ with highest rank $R(XY; A_n|\{A_i\}_{n-1})$:

  - Update its contributor list: $\{A_i\}_n \leftarrow \{A_i\}_{n-1} + A_n$
  - If $I(X;Y|\{A_i\}_n)$ is not significant (given the finite number $N$ of samples): remove edge $XY$
  - Else: Search for the next best contributor $A_{n+1}$ of edge $XY$ (if one exists with $I(X;Y;A_{n+1}|\{A_i\}_n) > 0$) and update the ranking order $R(XY; A_{n+1}|\{A_i\}_n)$

  **Until**: no more edges can be removed

- Step 2: *Confidence estimate and sign of retained edges*

  Once a first skeleton has been obtained using Step 1, the confidence on each retained edge can be estimated through an edge specific confidence ratio $C_{XY}$ based on the probability $P_{XY} \sim \exp(-NI(X;Y|\{A_i\}))$ to remove a directed edge $X \to Y$ from the graph $\mathcal{G}$ (S1 File),

  $$C_{XY} \quad = \quad \frac{P_{XY}}{\langle P_{XY}^{\text{rand}} \rangle} \tag{12}$$

  where $\langle P_{XY}^{\text{rand}} \rangle$ is the average of the probability to remove the $XY$ edge after randomly permutating the dataset for each variable. Hence, the lower $C_{XY}$, the higher the confidence on the $XY$ edge. We favor the confidence estimate $C_{XY}$ based on likelihood ratios (Eq. S21 in S1 File) to the alternative confidence estimate based on p-value, which corresponds to the probability that $P_{XY}^{\text{rand}} \leqslant P_{XY}$ over random permutations. Indeed, p-value estimates require much more random permutations than $C_{XY}$ estimates for strong edges with $NI(X;Y|\{A_i\}) \gg 1$, as virtually all random permutations correspond to $P_{XY}^{\text{rand}} > P_{XY}$ in that case, leading to under-estimated p-values $\simeq 0$.

  In addition, the sign of each retained edge, $X - Y$, is defined by the sign of the partial correlation coefficient, $\rho_{XY \cdot \boldsymbol{A}}$, between $X$ and $Y$ conditioned on its derived contributors $\boldsymbol{A} = \{A_i\}$ in Step 1, with positive edges corresponding to

positive partial correlations and negative edges corresponding to negative partial correlations, *i.e.* partial anti-correlations (S1 File).

- Step 3: *Probabilistic orientation and propagation of remaining edges*

  Given the skeleton obtained from Step 1, possibly filtered through Step 2, initially unspecified endpoint marks ($\circ$) can be established, as arrow tail ($-$) or head ($>$), following probabilistic orientation and propagation rules of unshielded triples $\langle X, Y, Z\rangle_{X\neq Y}$, S1 File (where $*$ below stands for any endpoint mark),

  **Repeat**: take the top $\langle X, Y, Z\rangle_{X\neq Y}$ with highest endmark orientation / propagation probability

  - If $I(X;Y;Z|\{A_i\}_n) < 0$ and $X*\!-\circ Z\circ\!-\!*Y$ or $X*\!\rightarrow Z\circ\!-\!*Y$, orient edge(s) to form a v-structure $X*\!\rightarrow Z\leftarrow\!*Y$
  - Else If $I(X;Y;Z|\{A_i\}_n) > 0$ and $X*\!\rightarrow Z\circ\!-\!\circ Y$ or $X*\!\rightarrow Z\circ\!\rightarrow Y$, Propagate second edge direction to form a non-v-structure $X*\!\rightarrow Z\rightarrow Y$

  **Until**: no additional endmark orientation / propagation probability $>1/2$

### Algorithmic performance on benchmark networks with latent variables

The performance of the information-theoretic method `miic` was tested on benchmark ancestral graphs with latent variables using partially observed real-life networks (*i.e.* considering some variables as hidden) as well as random networks generated with the causal modeling tool Tetrad V (`http://www.phil.cmu.edu/tetrad`). Reconstructed networks are compared to *partial ancestral graphs* (PAGs) [23], which are the representatives of the Markov equivalent class of all ancestral graphs consistent with the conditional independences in the available data. In practice, benchmark PAGs have been derived by hidding some variables in benchmark directed acyclic graphs (DAG) using the `dag2pag` function of the `pcalg` package with slight modifications [25, 26]. PAGs have been generated for an increasing fraction (0% to 20%) of randomly picked latent variables having a significant topological effect on the underlying network (*i.e.* excluding parentless vertices with a single child or vertices without child).

The results are evaluated in terms of skeleton Precision (or positive predictive value), $Prec = TP/(TP+FP)$, Recall or Sensitivity (true positive rate), $Rec = TP/(TP+FN)$, as well as F-score $= 2\times Prec\times Rec/(Prec+Rec)$ for increasing sample size from $N$=10 to 50,000 data points. We also define additional Precision, Recall and F-scores taking into account the edge endpoint marks of the predicted networks against the corresponding benchmark PAGs. This amounts to label as false positives, all true positive edges of the skeleton with different arrowhead endpoint marks (*i.e.* arrowhead ($>$) *versus* tail or undefined ($-/\circ$) endpoint marks) as the PAG reference, $TP_{\text{misorient}}$, leading to the orientation-dependent definitions $TP' = TP - TP_{\text{misorient}}$ and $FP' = FP + TP_{\text{misorient}}$ with the corresponding PAG Precision, Recall and F-scores taking into account arrowhead endpoint marks.

The alternative inference methods used for comparison with `miic` are the `FCI` algorithm [9] and its recent approximate variant `RFCI` [10] implemented in the `pcalg` package [25, 26]. The results obtained with `FCI` and `RFCI` are in fact very similar and we only present here comparisons with the more recent `RFCI` algorithm [10]. `RFCI`'s results are shown for an adjustable significance level $\alpha = 0.01$ and using the *stable*

implementation of the skeleton learning algorithm, as well as the *majority rule* for the orientation and propagation steps [27], which give overall the best results.

For each sample size ($N$=10 to 50,000) and fraction of hidden variables (0% to 20%), `miic` and `RFCI` inference methods have been tested on 20 combinations of hidden variables and 50 dataset replicates each. S1 Fig, S2 Fig and S3 Fig give the average results over these multiple combinations of latent variables and dataset replicates and compare the reconstructed networks including orientations (solid lines) or without orientation (*i.e.* skeleton, dashed lines) to the theoretical PAG (or its skeleton) of the benchmark network.

### Algorithmic performance on undirected benchmark networks

The performance of `miic` was also tested on non-causal benchmark networks reconstructed from Monte Carlo sampling of Ising-like interacting systems.

To this end, real-life causal networks, such as Alarm and Insurance, have been transformed into non-causal Ising-like networks (with binary spin variables $x_i = \pm 1$) by setting pairwise interacting parameters $k_{ij}$ between connected variables $X_i$ and $X_j$, so as to approximately reproduce the pairwise conditional mutual information $I(X_i; X_j | \boldsymbol{A}_{X_i X_j})$ of the original real-life causal network. This yields benchmark networks sharing approximately the same two-point direct correlations with the original causal networks but lacking causality, as the couplings $k_{ij}$ between spins are all symmetric by construction.

One million configurations of these Ising-like interacting systems have been generated using Monte Carlo sampling approach. It consists in flipping a fraction of the spins randomly and accepting each newly generated configuration with probability, $\min\big(1, \exp(-\Delta E_k)\big)$, where $\Delta E_k = E_{k+1} - E_k$, is the interacting energy difference between successive configurations, $E_k = -\sum_{i<j}^{\text{edges}} k_{ij} x_i x_j$. The fraction of spins randomly flipped ($\sim$10%) has been ajusted to ensure that about half of the newly generated configurations are accepted at each Monte Carlo iteration, in order to efficiently sample configuration space. This leads, however, to significant correlations between successive accepted configurations with a roughly exponential decay between $n$ distant samples, $C(n) \simeq C(0) \exp(-n/R) = C(0)\alpha^n$, where $C(n) = C(k-\ell) = \langle \sum_i \delta x_i^{(\ell)} \delta x_i^{(k)} \rangle$ is the average autocorrelation with lag between the $k$th and $\ell$th samples (with $n = k - \ell$), where $\delta x_i^{(k)} = x_i^{(k)} - \bar{x_i}$.

The effective number of independent samples $N^*_{\text{eff}}$ can then be estimated through the apparent increase of variance between the $N$ partially correlated samples as [67],

$$\begin{aligned} V_N &= \frac{1}{N^2} \sum_k \sum_\ell \langle \sum_i \delta x_i^{(k)} \delta x_i^{(\ell)} \rangle \\ &= \frac{1}{N^2} \sum_k \sum_\ell C(k-\ell) \\ &= \frac{1}{N} \left[ C(0) + 2\Big(1 - \frac{1}{N}\Big) C(1) + 2\Big(1 - \frac{2}{N}\Big) C(2) + \cdots + \frac{2}{N} C(N-1) \right] \end{aligned} \quad (13)$$

which leads for a first order Markov process with $C(n) = C(0)\alpha^n$ to,

$$\begin{aligned} V_N &= \frac{C(0)}{N} \left[ 1 + 2\Big(1 - \frac{1}{N}\Big)\alpha + 2\Big(1 - \frac{2}{N}\Big)\alpha^2 + \cdots + \frac{2}{N}\alpha^{N-1} \right] \\ &\simeq \frac{C(0)}{N} \frac{1+\alpha}{1-\alpha} = \frac{C(0)}{N^*_{\text{eff}}} \end{aligned} \quad (14)$$

yielding a smaller effective number of samples $N^*_{\text{eff}} < N$ for correlated datasets ($\alpha > 0$) as,

$$N^*_{\text{eff}} \quad = \quad N\frac{1-\alpha}{1+\alpha} \tag{15}$$

This estimate suggests to use $N^*_{\text{eff}}$, instead of $N$, to compute the finite size corrections of the `miic` approach, in order to correct for the correlations between successive samples generated through Monte Carlo sampling. Yet, as the presence of correlations between successive samples is *a priori* incompatible with the requirement of independent samples in the maximum likelihood framework, we have first assessed `miic` performance over the full range of possible effective sample size, *i.e.* $0 < N_{\text{eff}}/N \leqslant 1$, for $N = 1,000$ to $300,000$ successive samples from the one-million-long sample series.

The results are shown in S6 Fig and S6 Fig in terms of Precision, Recall, F-score and Fraction of (wrongly) directed edges for the Alarm-like and Insurance-like undirected networks.

The nearly exponential decay of the autocorrelation function for Alarm-like (S6 Fig, $R = 7.758$, $\alpha = 0.872$) and Insurance-like (S6 Fig, $R = 7.676$, $\alpha = 0.87$) undirected networks leads to very close values for the predicted effective number of samples for these graphs according to Eq. 15, $N^*_{\text{eff}}/N \simeq 0.068 - 0.069$.

Interestingly, we found that the F-score, which is a trade-off between optimizing Precision and Recall, reaches a maximum for all sample sizes ($N = 1,000$ to $300,000$) around the predicted effective number of samples, that is when $N_{\text{eff}}/N = N^*_{\text{eff}}/N \simeq 0.069$, see vertical dashed lines in F-score in S6 Fig and S6 Fig. We found also that the fraction of (wrongly) directed edges is close to zero at the predicted effective number of samples, $N^*_{\text{eff}}$, providing that it is not too small, *i.e.* $N^*_{\text{eff}} > 300$.

These results demonstrate that the theoretical estimate of $N^*_{\text{eff}}$, Eq. 15, yields the best compromise between over-fitting and under-fitting graphical models given the finite and partially correlated available datasets. They underline also `miic` accuracy to discard spurious causality in observational data, even from relatively small effective numbers of independent samples, *i.e.* $N^*_{\text{eff}} > 300$ in S6 Fig and S6 Fig.

# Supporting information

**S1 File. Supplementary text.** Contents: **1** Information-theoretic approach to network reconstruction ; **1.1** Signature of causality *versus* indirect contributions to information in graphs; **1.2** Finite size effect and most likely contributor score. **2** Algorithmic pipeline of the information-theoretic approach `miic`; **2.1** Algorithm 1: Learning skeleton taking into account latent variables; **2.2** Algorithm 2: Confidence estimation and sign of retained edges; **2.3** Algorithm 3: Probabilistic orientation and propagation of remaining edges. **3** Algorithmic implementation and tools; **3.1** `miic` R-package; **3.2** `miic` and `FCI` executables. **4** References for Supplementary Text.

**S1 Fig. Real-life Alarm network with hidden latent variables** [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall, F-score and computing time for PAG skeletons (dashed lines) and PAGs including orientations (solid lines). The results are given for the `miic` algorithm (warm colors) compared to the `RFCI` algorithm [10] (cold colors) for 0, 2, 4 and 6 latent variables out of the 37 nodes. Computation times in log scale show a linear scaling in the limit of

large datasets, $\tau_{\text{cpu}} \sim N^{0.9}$, for the `miic` algorithm, and a stronger nonlinear increase, $\tau_{\text{cpu}} \sim N^{1.5}$, with the `RFCI` algorithm.

**S2 Fig. Real-life Insurance network with hidden latent variables** [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall, F-score and computing time for PAG skeletons (dashed lines) and PAGs including orientations (solid lines). The results are given for the `miic` algorithm (warm colors) compared to the `RFCI` algorithm [10] (cold colors) for 0, 1, 2, and 4 latent variables out of the 27 nodes. Computation times in log scale show a linear scaling in the limit of large datasets, $\tau_{\text{cpu}} \sim N^{1.0}$, for the `miic` algorithm, and a stronger nonlinear increase, $\tau_{\text{cpu}} \sim N^{1.7}$, with the `RFCI` algorithm.

**S3 Fig. Real-life Barley network with hidden latent variables** [48 nodes, 84 links, 114,005 parameters, Average degree 3.5, Maximum in-degree 4]. Precision, Recall, F-score and computing time for PAG skeletons (dashed lines) and PAGs including orientations (solid lines). The results are given for the `miic` algorithm (warm colors) compared to the `RFCI` algorithm [10] (cold colors) for 0, 2, 4 and 7 latent variables out of the 48 nodes. Computation times in log scale show a nearly linear scaling in the limit of large datasets, $\tau_{\text{cpu}} \sim N^{1.1}$, for the `miic` algorithm, and a stronger nonlinear increase, $\tau_{\text{cpu}} \sim N^{2.3}$, with the `RFCI` algorithm.

**S4 Fig. Reconstruction of Fig. 1C network from simulated data.** `miic` and `RFCI` [9, 10] *versus* `3off2` [19] and `PC` [7, 8, 25] reconstructions of Fig. 1C network are performed from simulated data generated with Tetrad V, $N = 10 - 50,000$ samples. Precision, Recall and Fscore are given for skeleton (dashed lines) and PAG including orientations (solid lines).

**S5 Fig. Random benchmark networks of increasing size.** `miic` reconstruction of random networks of increasing size ($P = 10 - 500$ nodes) and fixed average degree 3 from $N = 1,000$ samples generated with Tetrad V. The average CPU time exhibits an optimal quadratic complexity in terms of network size, $\tau_{\text{cpu}} \sim P^2$ (solid bar), with only a small time increase when considering latent variables (orange) as compared to excluding them (red).

**S6 Fig. Alarm-like undirected network.** Precision, Recall, F-score, percentage of (wrongly) directed edges and decay of the autocorrelation function with lag between successive samples for $N = 1,000$ to $300,000$ consecutive partially correlated samples (with predicted effective number of independent samples in brackets). Vertical dashed lines correspond to the predicted effective number of independent samples $N^*_{\text{eff}}/N \simeq 0.068$, see Materials and Methods.

**S7 Fig. Insurance-like undirected network.** Precision, Recall, F-score, percentage of (wrongly) directed edges and decay of the autocorrelation function with lag between successive samples for $N = 1,000$ to $300,000$ consecutive partially correlated samples (with predicted effective number of independent samples in brackets). Vertical dashed lines correspond to the predicted effective number of independent samples $N^*_{\text{eff}}/N \simeq 0.069$, see Materials and Methods.

**S8 Fig. Edge confidence filtering on real-life Alarm network** [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall, F-score and computing time for network skeleton (dashed lines) and oriented network

CPDAG (solid lines) for a decreasing edge-specific confidence filtering, $C_{XY} = 1$ (no filtering) 0.01, 0.001 and 0.0001. For sample size > 100, confidence filtering of individual edges improves the precision (at the expense of recall) not only for the skeleton (dashed lines), as expected, but also for the oriented networks (solid lines). In addition, limited filtering, *i.e.* keeping edges with $C_{XY} < 10^{-3} - 10^{-2}$, tends to yield equivalent F-scores as unfiltered benchmark reconstructions.

**S9 Fig. Edge confidence filtering on real-life Insurance network** [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall, F-score and computing time for network skeleton (dashed lines) and oriented network CPDAG (solid lines) for a decreasing edge-specific confidence filtering, $C_{XY} = 1$ (no filtering) 0.01, 0.001 and 0.0001. For sample size > 100, confidence filtering of individual edges improves the precision (at the expense of recall) not only for the skeleton (dashed lines), as expected, but also for the oriented networks (solid lines). In addition, limited filtering, *i.e.* keeping edges with $C_{XY} < 10^{-3} - 10^{-2}$, tends to yield equivalent F-scores as unfiltered benchmark reconstructions.

**S10 Fig. Gene expression distribution in 3,934 single cells from mouse embryos**. Expression data on the 33 TFs are obtained from [24]. Percentage of samples with expressed genes (red) and non-expressed genes (gray).

**S11 Fig. Unfiltered network skeleton for hematopoiesis differentiation data.** Hematopoietic / endothelial gene expression data in 3,934 single cells from mouse embryos [24]. 7 out of 82 edges (8.5%) with $C_{XY} > 10^{-3}$ have been filtered in Fig. 2D (blue edges correspond to anti-correlations).

**S12 Fig. Expression alterations in** 807 **samples of breast tumor data from COSMIC database** [34]. Percentage of samples with normalized over-expression (red), normalized under-expression (blue) and unchanged normalized expression (gray) for each gene based on COSMIC.

**S13 Fig. Mutations in** 807 **samples of breast tumor data from COSMIC database** [34]. Percentage of mutated samples (red) for each gene.

**S14 Fig. Unfiltered network skeleton for breast tumor ploidy-mutation-expression data from COSMIC database** [34]. Due to the limited numbers of samples (N=807) and recurrent gene mutants (Figure 3-figure supplement 2), most gene mutations (yellow) are not confidently linked to any altered expression levels (green) and have been filtered in the high confidence network Fig. 3D ($C_{XY} < 10^{-3}$), with the notable exceptions of *TP53* and *RB1* mutations, which have a significant impact on gene expressions, Fig. 3D, see main text (blue edges correspond to anti-correlations).

**S15 Fig. Unfiltered network skeleton for ohnolog retention data in human.** Genomic data for the 20,415 human coding genes is provided in S1 Data. The only edge with confidence ratio $C_{XY} > 10^{-3}$ is RecDominance — ProteinComplex with $C_{XY} = 0.25$ (blue edges correspond to anti-correlations).

**S1 Software. Software and tools.** `miic` software is provided in two formats: an R-package to be used in the R environment, and `miic` and `FCI` executables, which were used for all benchmarks included in the paper.

**S1 Data. Dataset of human genomic properties.** This dataset contains all genomic data for the 20,415 human genes analyzed in Fig. 4.

**S1 Table. Effect of BIC/MDL *versus* NML criteria in applications.** Choosing the BIC/MDL instead of NML criterion in the three genetic network applications, Fig. 2, Fig. 3 & Fig. 4, leads to somewhat sparser reconstituted networks including 82% to 100% of initial edges, yet no additional edges (*i.e.* consistent with a lower recall), and 66% to 75% conserved edge orientations (*i.e.* identical –, →, ← and ↔ edges).

## Acknowledgments

We thank François Graner, Isabelle Guyon, Giulia Malaguti, Philippe Marcq, Leonid Mirny, Leila Perie, Guido Uguzzoni, Jean-Philippe Vert, Martin Weigt for stimulating discussions.

## References

1. Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. Nat Methods. 2016;13(4):310–318. doi:10.1038/nmeth.3773.
2. Meinshausen N, Hauser A, Mooij JM, Peters J, Versteeg P, Buhlmann P. Methods for causal inference from gene perturbation experiments and validation. Proc Natl Acad Sci USA. 2016;113(27):7361–7368. doi:10.1073/pnas.1510493113.
3. Heckerman D, Geiger D, Chickering DM. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Mach Learn. 1995;20(3):197–243. doi:10.1023/A:1022623210503.
4. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9(3):432–441. doi:10.1093/biostatistics/kxm045.
5. Jaynes ET. On the rationale of maximum-entropy methods. Proceedings of the IEEE. 1982;70(9):939–952. doi:10.1109/proc.1982.12425.
6. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proc Natl Acad Sci USA. 2005;102(21):7426–7431. doi:10.1073/pnas.0500334102.
7. Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. Social Science Computer Review. 1991;9:62–72. doi:10.1177/089443939100900106.
8. Pearl J, Verma T. A theory of inferred causation. In: In Knowledge Representation and Reasoning: Proc. of the Second Int. Conf.; 1991. p. 441–452.
9. Spirtes P, Meek C, Richardson T. An Algorithm for causal inference in the presence of latent variables and selection bias. In: Computation, Causation, and Discovery. Menlo Park, CA: AAAI Press; 1999. p. 211–252.
10. Colombo D, Maathuis MH, Kalisch M, Richardson TS. Learning high-dimensional directed acyclic graphs with latent and selection variables. Ann Statist. 2012;40(1):294–321. doi:10.1214/11-aos940.

11. Richardson T, Spirtes P. Ancestral graph Markov models. Ann Statist. 2002;30(4):962–1030. doi:10.1214/aos/1031689015.

12. Claassen T, Mooij J, Heskes T. Learning sparse causal models is not NP-hard. In: UAI 2013, Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence; 2013. p. 172–181.

13. Pearl J. Causality: models, reasoning and inference. 2nd ed. Cambridge University Press; 2009.

14. 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; 2015. p. 42–51.

15. McGill WJ. Multivariate information transmission. Trans of the IRE Professional Group on Information Theory (TIT). 1954;4:93–111. doi:10.1007/BF02289159.

16. Ting HK. On the Amount of Information. Theory Probab Appl. 1962;7(4):439–447. doi:10.1137/1107041.

17. Han TS. Multiple Mutual Informations and Multiple Interactions in Frequency Data. Information and Control. 1980;46(1):26–45. doi:10.1016/S0019-9958(80)90478-7.

18. Yeung RW. A new outlook on Shannon's information measures. IEEE transactions on information theory. 1991;37(3):466–474. doi:10.1109/18.79902.

19. Affeldt S, Verny L, Isambert H. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. BMC Bioinformatics. 2016;17(S2).

20. Cover TM, Thomas JA. Elements of Information Theory. 2nd ed. Wiley-Interscience; 2006.

21. Rebane G, Pearl J. The recovery of causal poly-trees from statistical data. Int J Approx Reasoning. 1988;2(3):341. doi:http://fmdb.cs.ucla.edu/Treports/870031.pdf.

22. Uda S, Saito TH, Kudo T, Kokaji T, Tsuchiya T, Kubota H, et al. Robustness and Compensation of Information Transmission of Signaling Pathways. Science. 2013;341(6145):558–561. doi:10.1126/science.1234511.

23. Zhang J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artif Intell. 2008;172(16-17):1873–1896. doi:10.1016/j.artint.2008.08.001.

24. Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol. 2015;33(3):269–276. doi:10.1038/nbt.3154.

25. Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P. Causal inference using graphical models with the R package pcalg. J Stat Softw. 2012;47(11):1–26. doi:10.18637/jss.v047.i11.

26. Kalisch M, Bühlmann P. Robustification of the PC-Algorithm for Directed Acyclic Graphs. J Comp Graph Stat. 2008;17(4):773–789. doi:stat.ethz.ch/ kalisch/papers/pcrobMay08.pdf.

27. Colombo D, Maathuis MH. Order-Independent Constraint-Based Causal Structure Learning. J Mach Learn Res. 2014;15:3741–3782. doi:http://jmlr.org/papers/v15/colombo14a.html.

28. Kontkanen P, Myllymäki P. A linear-time algorithm for computing the multinomial stochastic complexity. Inf Process Lett. 2007;103(6):227–233.

29. Roos T, Silander T, Kontkanen P, Myllymäki P. Bayesian network structure learning using factorized NML universal models. In: Proc. 2008 Information Theory and Applications Workshop (ITA-2008). IEEE Press; 2008.

30. Baron MH. Concise Review: early embryonic erythropoiesis: not so primitive after all. Stem Cells. 2013;31(5):849–856. doi:10.1002/stem.1342.

31. Ferreiros-Vidal I, Carroll T, Taylor B, Terry A, Liang Z, Bruno L, et al. Genome-wide identification of Ikaros targets elucidates its contribution to mouse B-cell lineage specification and pre-B-cell differentiation. Blood. 2013;121(10):1769–1782. doi:10.1182/blood-2012-08-450114.

32. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Anagnostopoulos A, et al. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Res. 2015;43(Database issue):D726–736. doi:10.1093/nar/gku967.

33. Pimanda JE, Ottersbach K, Knezevic K, Kinston S, Chan WY, Wilson NK, et al. Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. Proc Natl Acad Sci USA. 2007;104(45):17692–17697. doi:10.1073/pnas.0707045104.

34. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43(D1):D805–D811. doi:10.1093/nar/gku1075.

35. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45(10):1134–1140. doi:10.1038/ng.2760.

36. Kollareddy M, Dimitrova E, Vallabhaneni KC, Chan A, Le T, Chauhan KM, et al. Regulation of nucleotide metabolism by mutant p53 contributes to its gain-of-function activities. Nat Commun. 2015;6:7389. doi:10.1038/ncomms8389.

37. Swanton C, McGranahan N, Starrett GJ, Harris RS. APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. Cancer Discov. 2015;5(7):704–712. doi:10.1158/2159-8290.CD-15-0344.

38. Aylon Y, Oren M. p53: guardian of ploidy. Mol Oncol. 2011;5(4):315–323. doi:10.1016/j.molonc.2011.07.007.

39. Dewhurst SM, McGranahan N, Burrell RA, Rowan AJ, Gronroos E, Endesfelder D, et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. Cancer Discov. 2014;4(2):175–185. doi:10.1158/2159-8290.CD-13-0285.

40. Kuznetsova AY, Seget K, Moeller GK, de Pagter MS, de Roos JA, Durrbaum M, et al. Chromosomal instability, tolerance of mitotic errors and multidrug resistance are promoted by tetraploidization in human cells. Cell Cycle. 2015;14(17):2810–2820. doi:10.1080/15384101.2015.1068482.

41. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009;10(10):725–732. doi:10.1038/nrg2600.

42. Makino T, McLysaght A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci USA. 2010;107(20):9270. doi:10.1073/pnas.0914697107.

43. Singh PP, Affeldt S, Cascone I, Selimoglu R, Camonis J, Isambert H. On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. Cell Rep. 2012;2(5):1387–1398. doi:10.1016/j.celrep.2012.09.034.

44. Singh PP, Affeldt S, Malaguti G, Isambert H. Human dominant disease genes are enriched in paralogs originating from whole genome duplication. PLoS Comput Biol. 2014;10(7):e1003754. doi:10.1371/journal.pcbi.1003754.

45. Singh PP, Arora J, Isambert H. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. PLoS Comput Biol. 2015;11(7):e1004394. doi:10.1371/journal.pcbi.1004394.

46. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2009;19(2):327–335. doi:10.1101/gr.073585.107.

47. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet Genome Res. 2006;115(3-4):205–214. doi:10.1159/000095916.

48. Malaguti G, Singh PP, Isambert H. On the retention of gene duplicates prone to dominant deleterious mutations. Theor Popul Biol. 2014;93:38–51. doi:10.1016/j.tpb.2014.01.004.

49. Ganem NJ, Storchova Z, Pellman D. Tetraploidy, aneuploidy and cancer. Curr Opin Genet Dev. 2007;17(2):157–162. doi:10.1016/j.gde.2007.02.011.

50. Popova T, Manie E, Rieunier G, Caux-Moncoutier V, Tirapo C, Dubois T, et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. Cancer Res. 2012;72(21):5454–5462. doi:10.1158/0008-5472.CAN-12-1470.

51. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351(27):2817–2826. doi:10.1056/NEJMoa041588.

52. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530–536. doi:10.1038/415530a.

53. Buyse M, Loi S, Van't Veer L, Viale G, Delorenzi M, Glas AM, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst. 2006;98(17):1183–1192. doi:10.1093/jnci/djj329.

54. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–674. doi:10.1016/j.cell.2011.02.013.

55. Tian S, Roepman P, van't Veer LJ, Bernards R, De Snoo F, Glas AM. Biological functions of the genes in the mammaprint breast cancer profile reflect the hallmarks of cancer. Biomarker insights. 2010;5:129. doi:10.4137/BMI.S6184.

56. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70. doi:10.1038/nature11412.

57. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database (Oxford). 2010;2010:baq020. doi:10.1093/database/baq020.

58. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, et al. Natural selection on genes that underlie human disease susceptibility. Curr Biol. 2008;18(12):883–889. doi:10.1016/j.cub.2008.04.074.

59. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39(Web Server issue):29–37. doi:10.1093/nar/gkr367.

60. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40(Database issue):290–301. doi:10.1093/nar/gkr1065.

61. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database–2009 update. Nucleic Acids Res. 2009;37(Database issue):D767–772. doi:10.1093/nar/gkn892.

62. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. 2010;38(Database issue):497–501. doi:10.1093/nar/gkm936.

63. Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, et al. A census of human soluble protein complexes. Cell. 2012;150(5):1068–1081. doi:10.1016/j.cell.2012.08.011.

64. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics. 2010;8(1):77–80. doi:10.1016/S1672-0229(10)60008-3.

65. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA. 2004;101(16):6062–6067. doi:10.1073/pnas.0400782101.

66. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol. 2009;10(11):R130. doi:10.1186/gb-2009-10-11-r130.

67. Jones RH. Estimating the Variance of Time Averages. J Appl Meteor. 1975;14(2):159–163. doi:10.1175/1520-0450(1975)014¡0159:etvota¿2.0.co;2.

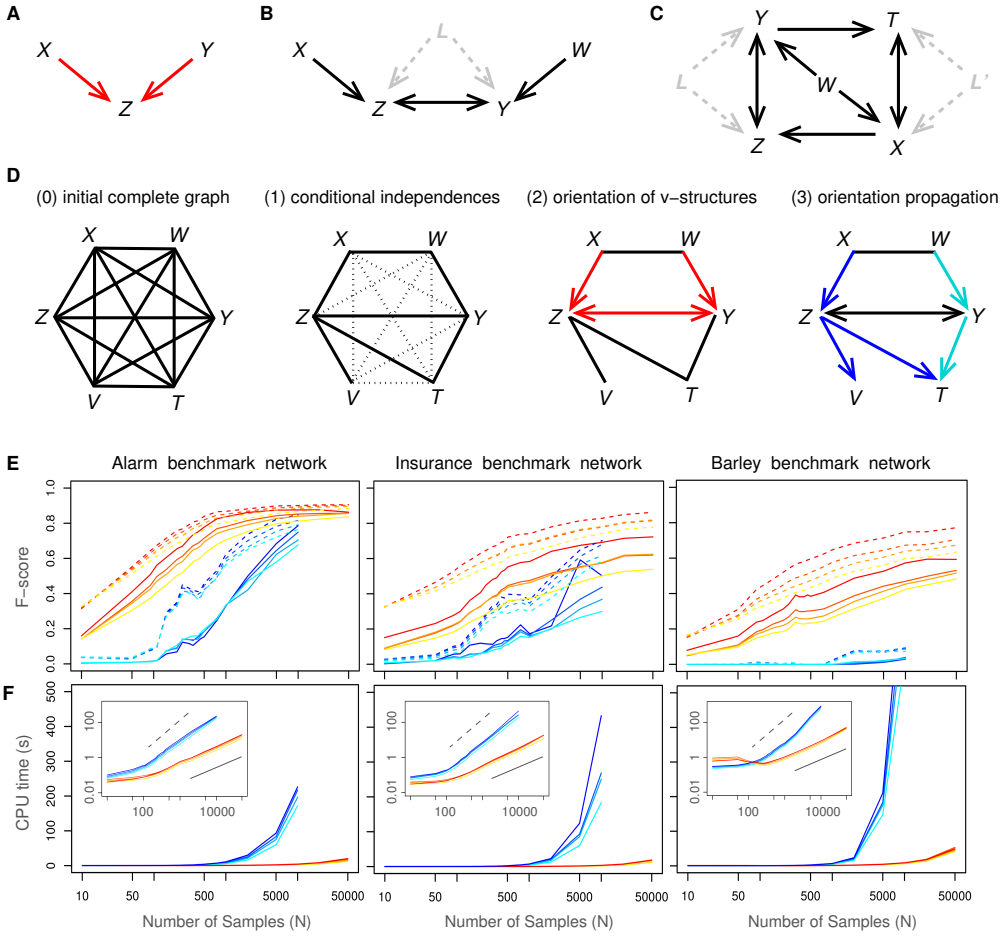



**Fig 1.** Learning causal networks with latent variables. (**A**) A v-structure. (**B**) Bidirected edges in collider paths indicate the presence of latent common cause(s), $L$, unobserved in the dataset. (**C**) Conditional independence in the presence of latent variables requires to be conditioned on non-adjacent variables, in general [9, 10], such as for the pair $\{Z, T\}$ which needs to be conditioned on $X$, $Y$ and non-adjacent $W$, $I(Z;T|X,Y,W) = 0$, as one cannot condition on the unobserved latent variables, $L$ or $L'$, *e.g.* $I(Z;T|X,L) = 0$ or $I(Z;T|Y,L') = 0$. (**D**) Outline of the successive steps of constraint-based approaches (see also Algorithm steps in Materials and Methods). (**E**) F-score (harmonic mean of Precision and Recall, S1 Fig, S2 Fig and S3 Fig) of `miic` algorithm (warm colors) for 0%, 5%, 10% and 20% of latent variables (top to bottom curves), compared to the `RFCI` algorithm [10] (cold colors) on benchmark networks of increasing complexity disregarding (dashed lines) or including (solid lines) edge orientations: Alarm [37 nodes, avg. deg. 2.5, 509 parameters], Insurance [27 nodes, avg. deg. 3.9, 984 parameters] and Barley [48 nodes, avg. deg. 3.5, 114,005 parameters]. (**F**) Computation times of `miic` (warm colors) compared to `RFCI` (cold colors). Inserts: computation times in log scale showing a linear scaling (solid bar) in the limit of large datasets, $\tau_{\text{cpu}} \sim N^{1\pm0.1}$, with `miic`, and a close to quadratic scaling (dashed bar), $\tau_{\text{cpu}} \sim N^{1.8\pm0.3}$, with `RFCI`.

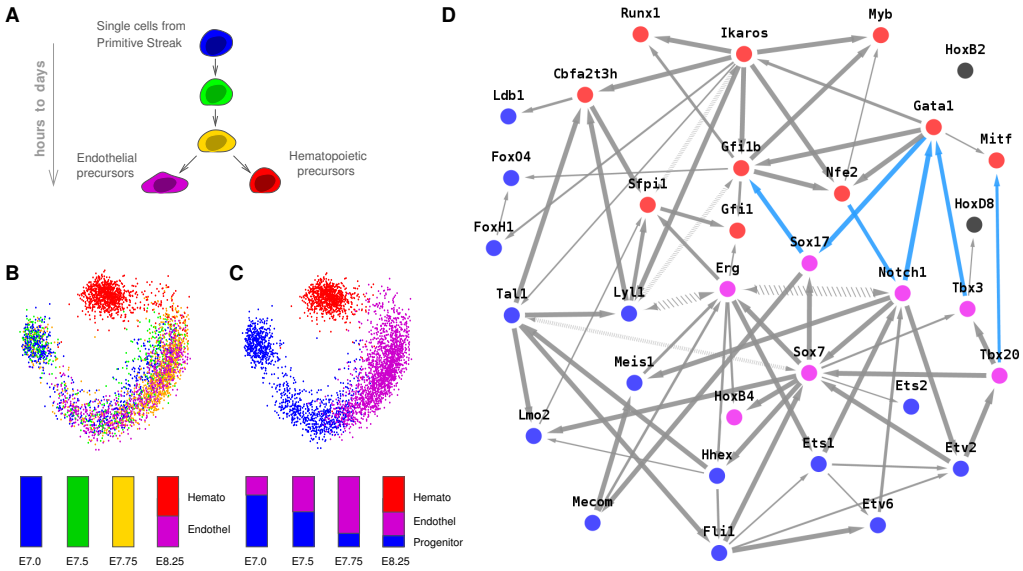



**Fig 2.** Network reconstruction at cellular level. (**A**) Hematopoietic / endothelial differentiation in single cells from mouse embryos [24]. (**B**) Principal component analysis and (**C**) K-means clustering of gene expression data [24] with histograms showing the relative proportions of cell populations at each data point (E7.0 to E8.25). (**D**) Hematopoietic / endothelial differentiation regulatory network between hematopoietic specific (red), endothelial (violet), common (blue) and unclassified (gray) TFs. Graph predicted with `miic` R-package and visualized using cytoscape (blue edges correspond to repressions).

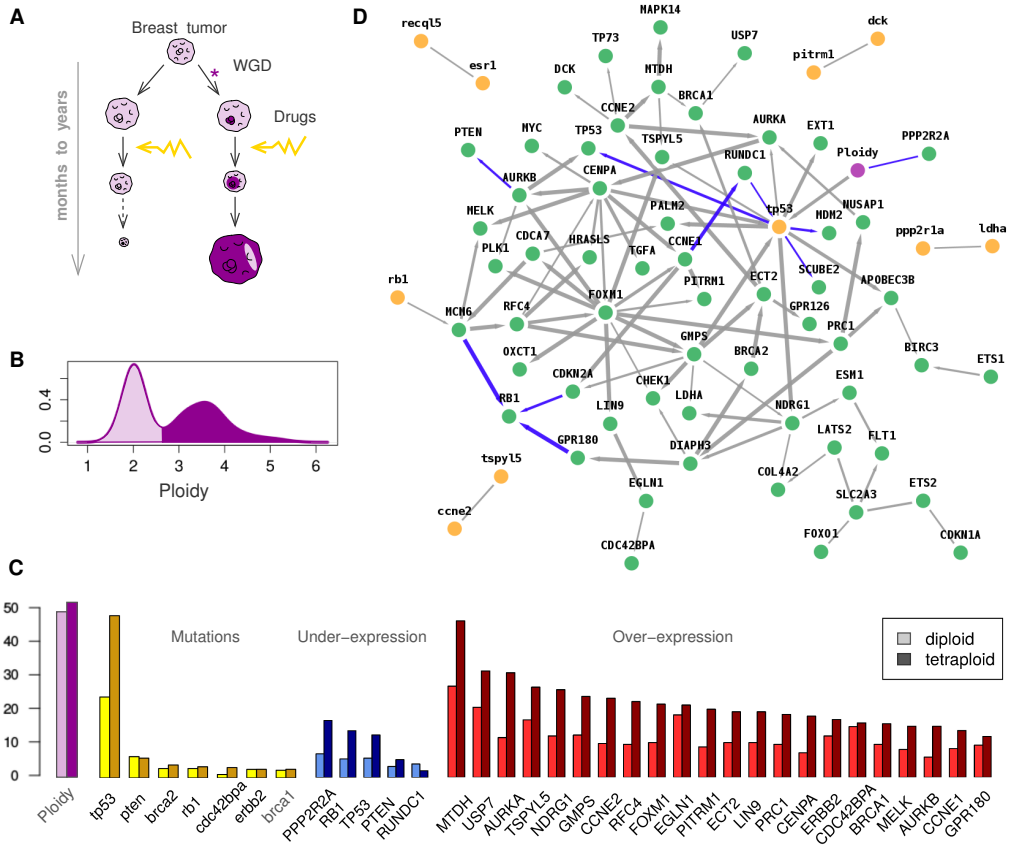



**Fig 3.** Network reconstruction at tissue level. (**A**) Tumor development and drug resistance in the presence of tetraploid tumor cells following whole genome duplication (WGD). (**B**) Ploidy distribution in the 807 tumor samples and (**C**) genomic alterations: ploidy, mutations, normalized under-expression and over-expression changes from COSMIC database [34]. (**D**) Genomic alteration network obtained between average ploidy (violet), gene mutations (yellow, lower case) and under- or over-expressions (green, upper case). Graph predicted with `miic` R-package and visualized using cytoscape (blue edges correspond to repressions).

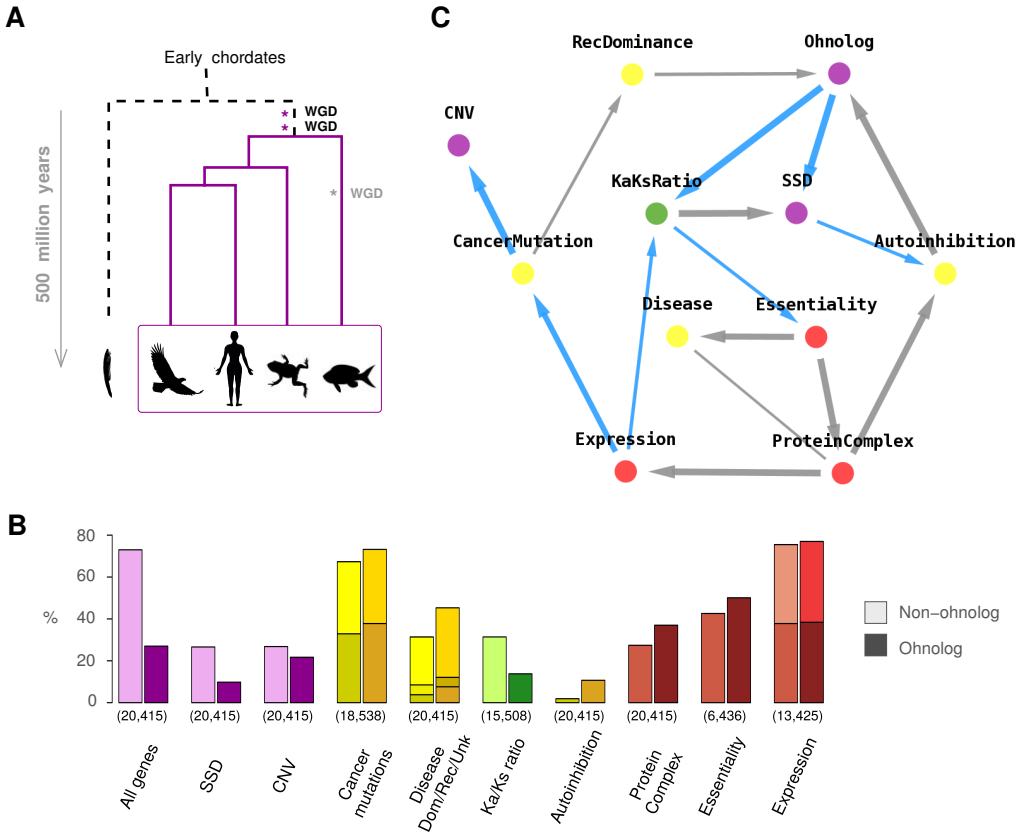



**Fig 4.** Network reconstruction at organismal and phylogenetic levels. (**A**) Two rounds of whole genome duplication (WGD) have led to the evolutionary radiation of vertebrates (and similarly with a third 300-MY-old WGD in teleost fish). (**B**) Biased distributions of genomic properties within 'non-ohnolog' and 'ohnolog' genes retained from WGDs in early vertebrates [45]. Numbers in brackets indicate the numbers of genes for which each property is identified, Materials and Methods and S1 Data. (**C**) Genomic property network of human genes, see main text. Graph predicted with `miic` R-package and visualized using cytoscape (blue edges correspond to repressions).

TROISIÈME PARTIE

# Utilisation de MIIC pour l'analyse de données réelles

CHAPITRE 8

# RÉGULATION TRANSCRIPTIONNELLE DE L'HÉMATOPOÏÈSE

## 8.1 Généralités

L'hématopoïèse est le processus au cours duquel des cellules pluripotentes (c'est-à-dire possédant la capacité de se *différencier* en plusieurs types cellulaires ainsi que de s'auto-renouveler) - dans ce cas appelées cellules souches hématopoïétiques - se différencient pour donner l'ensemble des lignées des cellules du sang. Ces lignées comprennent d'une part les érythrocytes (globules rouges) chargés du transport de l'oxygène d'autre part les cellules du système immunitaire (macrophages, lymphocytes T...) et enfin les plaquettes qui interviennent dans la coagulation. Le siège de ce processus varie en fonction du temps chez l'embryon (sac vitellin, aorte dorsale, foie, rate), mais chez l'adulte, l'hématopoïèse intervient majoritairement dans la moelle osseuse des os plats (os des hanches par exemple, appelé os iliaque).

Les régulations transcriptionnelles, consistant à moduler l'activité d'un gène en agissant sur la transcription de l'ADN en ARN, sont au cœur de ce mécanisme de différentiation. Certaines protéines, appelées facteurs de transcription (TF), sont dédiées à cette fonction. Les TF sont capables de se lier à l'ADN (avec ou sans partenaire(s)), ce qui a pour effet soit de stimuler la transcription du gène cible (*enhancers*), soit de la réprimer (*silencers*). Étant donnée l'implication majeure des dérèglements du réseau d'interactions transcriptionnelles dans des maladies telles que la leucémie, de nombreuses études se sont proposées d'étudier ces interactions à la fois en situation normale et pathologique ([44], [19], [23]).

Dans notre cas, nous avons tout d'abord utilisé le jeu de données d'une étude sur l'hématopoïèse chez l'adulte pour tenter de reconstituer les interactions entre 18 facteurs de transcription ayant un rôle clé dans ce processus [43]. Le caractère particulièrement bien étudié de l'hématopoïèse adulte nous a permis de "mesurer" la capacité de notre algo-

rithme `miic` à retrouver les liens pour lesquels une preuve expérimentale a été publiée [2].
Dans un second temps, nous avons voulu explorer un champ beaucoup moins étudié de l'hématopoïèse embryonnaire : la différenciation extra-embryonnaire des premiers globules rouges propres de l'embryon, appelée érythropoïèse primitive [4].

## 8.2 Hématopoïèse adulte

Comme expliqué précédemment, la reconstitution de ce réseau a été effectué avec comme objectif principal de tester les performances de `miic` dans un contexte de données biologiques réelles, c'est pourquoi la recherche de variables latentes n'a pas été utilisée dans ce cas précis.
Le jeu de données utilisé a été publié par Moignard et al. en 2013 [43]. Il est constitué de l'expression discrétisée de 18 facteurs de transcription possédant un rôle important dans le processus de différentiation des cellules hématopoïétique, mesurée dans 597 cellules uniques de 5 sous types de progéniteurs hématopoïétiques différents (fig.8.1). Nous avons

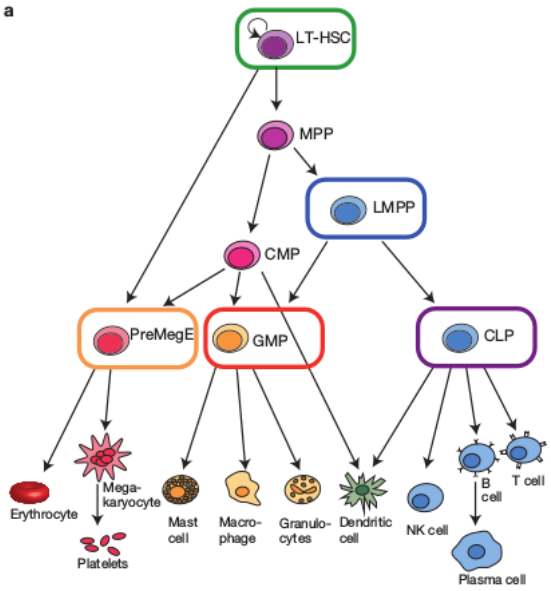



FIGURE 8.1 – **Arbre modélisant la différentiation des cellules hématopoïétiques chez l'adulte** – Les cadres colorés indiquent les progéniteurs présents dans le dataset. Figure extraite de *Moignard et al., Nat. Cell. Biol, 2013* [43]

reconstruit ce réseau en utilisant `3off2` et avons comparé le résultat obtenu à d'autres méthodes de reconstruction de réseaux classiquement utilisées pour l'étude de ce genre de problématique.

Le réseau reconstitué par `3off2` présente 41 liens, parmi lesquels la triple régulation entre les facteurs de transcription *Gata2*, *Gfi1*, *Gfi1b* confirmée expérimentalement dans l'étude originale ([43]). L'exploration de la littérature a permis de dégager en tout 11 relations transcriptionnelles confirmées expérimentalement ; nous nous en sommes servis comme indicateur des performances de chacune des méthodes de reconstruction utilisées sur ce jeu de données. Les résultats sont résumés dans le tableau 8.2. Notre méthode est

| 11 known Regulatory interactions | References | 3off2 *NML* | PC $\alpha = 10^{-1}$ | PC $\alpha = 10^{-2}$ | MMHC *BDe* | MMHC *BIC* | Bayes hc *BDe* | Bayes hc *BIC* | Aracne $\epsilon = 0$ |
|---|---|---|---|---|---|---|---|---|---|
| Gata2 → Gfi1b | [36] | → | ⇾ | – | ↛ | ↛ | → | ↛ | ↛ |
| Gfi1 → Gata2 | [36] | → | → | – | → | ⇾ | → | ⇾ | – |
| Gfi1b ⇆ Gfi1 | [36] | ⇾ | ⇾ | – | ⇾ | ⇾ | ⇾ | ⇾ | – |
| Gfi1 → PU.1 | [37] | → | → | ↛ | ↛ | ↛ | → | → | – |
| Lyl1 → Gfi1 | [38] | → | ⇾ | ↛ | ↛ | ↛ | → | ⇾ | – |
| Ldb1 → Meis1 | [39] | → | ↛ | ↛ | ↛ | ↛ | ⇾ | ↛ | ↛ |
| Ldb1 → Lyl1 | [39] | → | ↛ | ↛ | ↛ | ↛ | ↛ | ↛ | ↛ |
| Erg → Lyl1 | [40] | → | ⇾ | – | → | → | → | ⇾ | – |
| Gata2 → Scl | [40] | ⇾ | → | – | → | → | → | → | – |
| Gfi1b → Meis1 | [41] | ⇾ | ⇾ | – | → | → | → | → | – |
| Gata1 → Gata2 | [42] | ⇾ | ⇾ | – | → | → | → | → | – |
| Correct edges (out of **11**) | (→/⇾/–) | **11** | **9** | **7** | **6** | **6** | **10** | **8** | **8** |
| - Correct orientations | (→) | 7 | 3 | 0 | 5 | 4 | 8 | 4 | 0 |
| - Mis/non-orientations | (⇾/ –) | 4 | 6 | 7 | 1 | 2 | 2 | 4 | 8 |
| Missing links | (↛) | **0** | **2** | **4** | **5** | **5** | **1** | **3** | **3** |

FIGURE 8.2 – **Résumé de la capacité des différents algorithmes à retrouver les liens prouvés expérimentalement** – Figure extraite de *Affeldt et al., BMC Bioinformatics, 2016* [2]. Pour plus de détails, voir 4.3

la seule à inférer les 11 relations transcriptionnelles références, et retrouve une causalité telle qu'énoncée dans la littérature pour 7 d'entre elles. On note également les bonnes performances de la méthode hill-climbing en utilisant Bayesian Dirichlet equivalent (BDE) comme fonction de score, qui infère 10 des 11 interactions transcriptionnelles et oriente correctement 8 d'entre elles.

## 8.3 Hématopoïèse embryonnaire

Durant le développement de l'embryon des mammifères, l'hématopoïèse se développe en plusieurs vagues successives avec différents sites clés pour chacune de ces vagues. La première activité de différentiation hématopoïétique de l'embryon apparaît au niveau de la membrane extra-embryonnaire appelée sac vitellin [4] et a lieu entre les jours 7.25 et 8.75 du développement de l'embryon de la souris. C'est à ce moment que sont différenciés les premiers érythrocytes propres à l'embryon grâce à un mécanisme relativement peu connu, notamment en raison de la difficulté d'étudier les embryons chez les mammifères. Malgré cela il a été démontré que cette différentiation est très différente de celle de de l'adulte, puisque ces cellules hématopoïétiques primitives ne se différencient pas à partir de cellules souches hématopoïétiques, mais émergent grâce à un processus décrit comme linéaire, aboutissant à l'émergence au sein du sac vitellin de structures appelées *blood islands*. Ces structures sont formées à partir du mésoderme (l'un des 3 feuillets primitifs de l'embryon) qui, selon l'un des modèles de différentiation établit, se différencie en hémangioblaste sous l'influence du facteur de transcription *Etv2*, puis en endothélium (c'est ainsi qu'on appelle la couche interne des vaisseaux sanguins) hémogénique (c'est-à-dire ayant conservé un potentiel hématopoïétique) grâce à *Tal1* (aussi connu sous le nom de *SCL*) (fig. 8.3) [26]. L'endothélium hémogénique présente alors toutes les caractéristiques

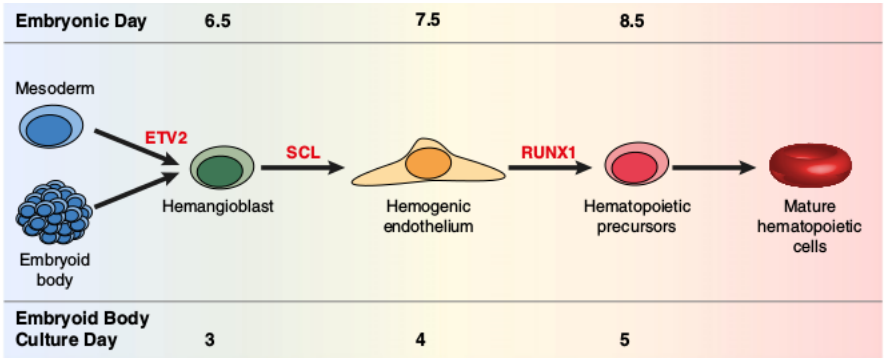



FIGURE 8.3 – **Processus de différentiation des précurseurs hématopoïétiques chez l'embryon** – tiré de *Moignard et al, Blood Cells, Mol. and Dis.,2013*

des cellules endothéliales, mais possède également toujours un potentiel hématopoïétique qui est exprimé au moment de la transition endothélio-hématopoïétique (*endothelial-to-haematopoietic transition*, EHT) d'où émerge les premières cellules sanguines et donc les *blood islands*[21], présentant à leur périphérie les cellules endothéliales primitives et les

cellules sanguines primitives au centre .

Afin d'étudier plus avant cette étape essentielle du développement de l'embryon, l'équipe de Berthold Göttgens a isolé 3934 cellules d'embryon de souris, à 4 moments différents du développement de celui-ci : jour 7.0, 7.5, 7.75 et 8.25 ; c'est à dire l'intervalle de temps au cours duquel ces premières cellules hématopoïétiques apparaissent. Le dernier jour de prélèvement a été séparé en deux pools : les cellules avec un potentiel endothélial ($Flk1^{+}/Runx1^{-}$) d'une part, et les cellules avec un potentiel hématopoïétique ($Flk1^{+}/Runx1^{+}$) d'autre part (fig. 8.4A). Les auteurs de l'étude ont ensuite quantifié l'expression de 33 facteurs de transcription ayant une implication documentée dans la différentiation endothéliale et/ou hématopoïétique, grâce à une technique appelée *microfluidic qRT-PCR* [25].
Ce type de PCR consiste à associer l'amplification de l'ADN de la PCR à l'action d'un fluorophore. Le nombre de cycles de PCR nécessaires pour la détection de chacune des longueurs d'ondes associées aux différentes séquences amplifiées (
$DeltaCt$) permet de connaître une approximation de la quantité initiale d'ARNm présents dans la cellule. Si, au bout d'un nombre de cycles préalablement défini, la fluorescence n'est pas observée, on déduit que le gène n'est pas exprimé de façon significative dans la cellule analysée. Moignard et al. ont utilisé cette dernière propriété pour rendre binaires les données d'expression, de sorte que chaque gène dans une cellule donnée est actif ($= 1$) ou inactif ($= 0$).

### 8.3.1 Analyse des données d'expression sur la totalité des cellules

Nous avons tout d'abord effectué la caractérisation des fonctions des 33 facteurs de transcription, utilisant pour cela la littérature existante sur des expériences de *knock-out* (KO - consiste à éteindre artificiellement et de façon définitive l'expression d'un gène ciblé) recensées au sein de la *Mouse Genome Database*[8], permettant ainsi de les séparer en trois groupes fonctionnels :

- Gènes Hématopoïétiques : Ce groupe rassemble les TFs pour lesquels nous n'avons trouvé qu'une fonction hématopoïétique, sans fonction endothéliale : Eto2, Sfpi1/PU.1, Runx1, Nfe2, Myb, Mitf, Ikaros, Gfi1b, Gfi1, Gata1
- Gènes Endothéliaux : A l'inverse, nous n'avons trouvé que des fonctions endothéliales pour ces gènes, sans fonction hématopoïétique associée : Ets2, Erg, Tbx3,

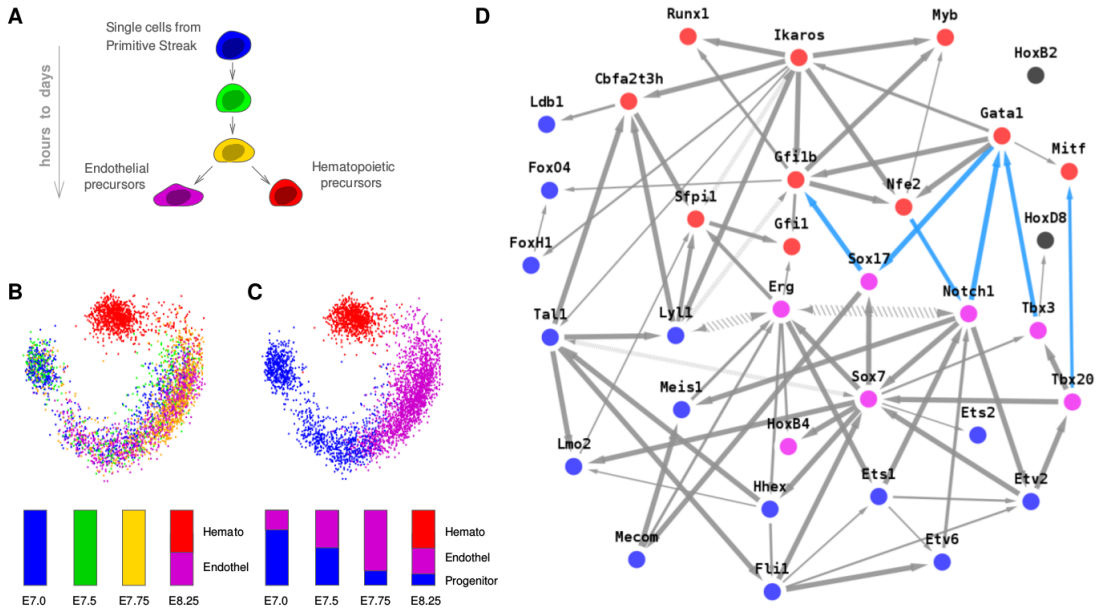



FIGURE 8.4 – **Reconstitution de réseau transcriptionnel** (**A**) Différentiation hématopoïétique/endothéliale de cellules uniques provenant d'embryons de souris. (**B**) Analyse en composante principale (représentation en 3 dimensions) et clustering (K-means) (**C**) des données d'expression de gènes. (**D**) Réseau de régulation de la différentiation hématopoïétique/endothéliale entre les facteurs de transcription impliqués dans la différentiation : gènes endothéliaux (violet), hématopoïétiques (rouge) et communs aux deux processus (bleu). Les liens bleus correspondent à des inhibitions.

Tbx20, Sox7, Sox17, Notch1, HoxB4

- Gènes non spécifiques : Pour les TFs de ce groupe nous avons trouvé une implication à la fois dans les différentiations endothéliale et hématopoïétique : Fli1, Etv6, Etv2, Ets1, Tal1, Meis1, Mecom, Lyl1, Lmo2, Ldb1, Hhex, FoxO4, FoxH1

Une fois la classification de chacun des facteurs de transcription retrouvée, nous avons pu nous intéresser au réseau reconstitué en utilisant l'ensemble des cellules, présentant plusieurs résultats intéressants (fig.8.4D).

En premier lieu, on note les répressions entre les gènes des deux groupes endothéliaux (violet) et hématopoïétiques (rouge) : sur les 8 interactions prédites entre ces deux groupes de facteurs de transcription, 6 sont des répressions. C'est un résultat attendu, étant données les fonctions de différentiation *a priori* opposées de ces deux groupes.

En second lieu, on note la présence de liens bi-dirigés, indiquant la présence probable d'une cause commune (ici un régulateur commun) non observée, ayant un impact sur la

transcription des deux gènes concernés. Les variables latentes prédites sur le réseau reconstruit semblent cohérentes avec la littérature ; notamment concernant la double orientation entre les gènes *Erg* et *Lyl1*, qui sont tous deux régulés par *Gata2* ([31]), non observé ici.
Enfin, un certain nombre des liens entre les différents facteurs de transcription ont été retrouvés dans la littérature, renforçant la cohérence de l'ensemble du réseau reconstitué. C'est le cas d'*Ikaros*, dont les interactions avec certains de ses partenaires sont évoquées dans [10], ou des interactions *Erg*, *Sox7* et *HoxB4*.

Une fois la cohérence du réseau global établie, l'objectif a été de voir s'il était possible de définir de nouvelles populations cellulaires qui présenteraient les caractéristiques des trois populations présentées dans le modèle de différentiation. Pour ce faire, nous avons utilisé une Analyse en Composante Principale (APC), permettant de visualiser trois sous populations distinctes, dont les frontières ont été définies grâce à l'algorithme des `k-moyens` [16] (fig. 8.4 BC).

Une fois ces 3 sous-populations isolées, les réseaux de régulation ainsi que les proportions d'expression des gènes de chacune d'entre elles ont été réalisés, afin de chercher une correspondance éventuelle avec les hémangioblastes, l'endothélium hémogénique et enfin les cellules hématopoïétiques primitives, conformément au modèle proposé (fig 8.3).

### 8.3.2 Sous-population *indifférenciée*

Ce cluster (en bleu sur la figure 8.4C) est majoritairement constitué de cellules prélevées aux jours 7.0 et 7.5 du développement de l'embryon, mais comprend également en proportion significative des cellules provenant des derniers moment d'extraction de l'expérience. Il apparaît que cette population tend à disparaître durant les stades plus avancés de l'expérience laissant progressivement la place à une autre population (en violet sur la figure 8.4C). Parmi ce groupe indifférencié, une majorité de cellules exprime certains facteurs de transcription non spécifiques (*FoxH1, FoxO4, Ldb1, Etv6* ou encore *Meis1*) mais également des facteurs à fonction plus endothéliale tels que *Ets2, Tbx3 et Tbx20*, tandis que les gènes du groupe à fonction plus hématopoïétique sont moins exprimés, à l'exception de *Runx1* ($\backsim 60\%$ des cellules) (fig.8.5A).

Il est en revanche intéressant d'observer que le réseau reconstitué sur cette sous-population, prédit un rôle central pour deux facteurs de transcription : *Tal1* et *Etv2* (fig.8.5B). Comme décrit plus haut, il a été démontré que *Etv2* joue un rôle important pour la pro-

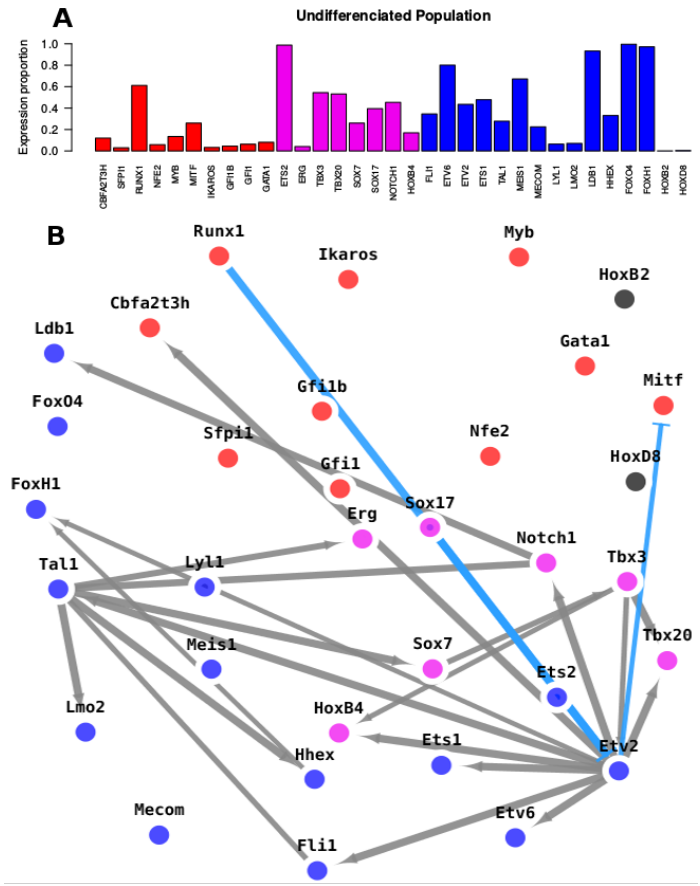



FIGURE 8.5 – **Caractérisation de la sous-population *indifférenciée*** (**A**) Proportion de cellules exprimant chacun des 33 facteurs de transcription, répartis en 3 groupes : hématopoïétiques (en rouge), endothéliaux (en violet) et communs (en bleu). (**B**) Réseau de régulation reconstitué sur les cellules appartenant au cluster indifférencié, en bleu sur la figure 8.4C

motion de la différentiation du mésoderme *Flk1*$^+$ en hémangioblaste tandis que *Tal1* a un rôle important dans le passage du stade d'hémangioblaste au stade d'endothélium hémogénique [26]. Ces caractéristiques pourraient donc correspondre à une population d'hémangioblastes hétérogènes, c'est-à-dire se trouvant à différents stades de leur différentiation :

- certaines cellules, effectuant la transition du stade mésodermique vers le stade d'hémangioblaste (et donc majoritairement sous l'influence d'*Etv2*,
- les autres cellules plus "précoces", qui entament quant à elles l'étape suivante du modèle de différentiation, vers l'endothélium hémogénique (régi par *Tal1*/*SCL*).

Le réseau reconstitué montre également une régulation positive de *Tal1* par *Etv2*, ce qui est cohérent avec la temporalité d'activation de ces deux gènes évoquée plus haut et déjà publié ([26]). Pour finir, la diminution relativement progressive de la proportion de cellules appartenant à cette population indifférenciée au cours du temps suggère une transition désynchronisées des cellules qui la composent au profit d'un autre groupe de cellules.

### 8.3.3 Précurseurs *endothéliaux*

Ce second cluster (en violet, fig.8.4C) est plus largement constitué de cellules prélevées aux jours 7.75 et 8.25 du développement, bien qu'il soit également présent aux premiers stades de l'expérience. La proportion de cellules lui appartenant augmente donc au cours du temps, suggérant un remplacement progressif de la population d'hémangioblastes décrits précédemment. Sur le plan de l'expression des facteurs de transcription, les caractéristiques présentées sont relativement différentes de la population indifférenciée. En effet, les cellules lui appartenant sont une majorité à exprimer indifféremment les facteurs de transcription associés aux trois groupes définis précédemment (fig.8.6A).
Le caractère non-spécifique de ces profils d'expression de cette sous-population cellulaire semble se confirmer lorsqu'on s'intéresse au réseau associé, qui prédit à la fois des interactions à l'intérieur des groupes de facteurs de transcription endothéliaux, hématopoïétiques et non-spécifiques (fig.8.6B), mais également de nombreuses interactions entre des facteurs de transcriptions appartenant à des groupes fonctionnels différents. On peut citer par exemple la présence de répressions entre les gènes endothéliaux tels que *Sox17* ou *Notch1* et des gènes à fonction majoritairement hématopoïétiques tels que *Ikaros, Gfi1b, Myb* ou *Gata1*.
Ces observations démontrent une certaine hétérogénéité de la population isolée et suggèrent la coexistence de deux potentiels de différentiation distincts : d'une part un destin

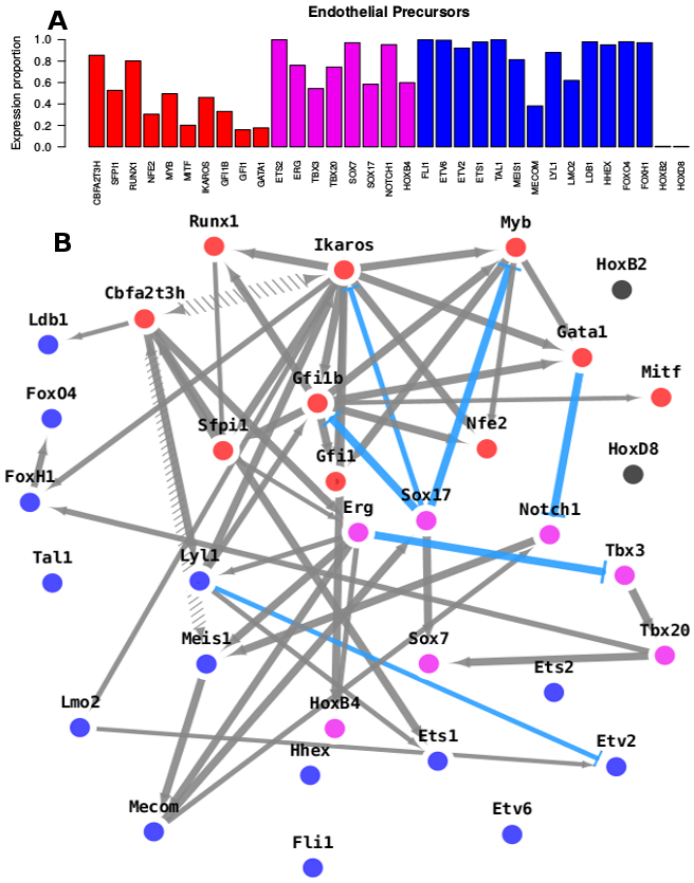



FIGURE 8.6 – **Caractérisation de la sous-population *endothéliale*** (**A**) Proportion de cellules exprimant chacun des 33 facteurs de transcription, répartis en 3 groupes : hématopoïétiques (en rouge), endothéliaux (en violet) et communs (en bleu). (**B**) Réseau de régulation reconstitué sur les cellules appartenant au cluster endothéliale, en violet sur la figure 8.4C

endothélial et d'autre part une capacité de différentiation hématopoïétique, une dualité propre à l'endothélium hémogénique. Il faut en revanche noter que la littérature indique un rôle clé de *Runx1* dans le processus de différentiation hématopoïétique, rôle qui semble occupé par *Ikaros* dans le réseau reconstruit, puisqu'il active 7 facteurs de transcription hématopoïétiques (dont *Runx1*) sur les 9 possibles. Il est possible que l'action essentielle de *Runx1* n'intervienne que sur un court laps de temps, entre les jours 7.75 et 8.25, expliquant l'absence de cellules présentant cette caractéristique dans le jeu de données utilisé.

### 8.3.4 Précurseurs *hématopoïétiques*

La dernière sous-population identifiée (en rouge sur la figure 8.4C) est quasi exclusivement constituée de cellules *Runx1*$^+$ prélevées au jour 8.25 de l'expérience. Les données d'expression représentées sur la figure 8.7A montre une population très différente des deux populations présentées précédemment. En effet, 100% de ces cellules expriment 5 facteurs de transcription hématopoïétiques sur 9, tandis qu'elles ne sont qu'une très faible minorité à exprimer des gènes à fonction endothéliale ($\leq 5\%$). Cette dernière caractéristique suggère que l'absence d'expression des facteurs de transcription endothéliaux est nécessaire pour que les cellules de l'endothélium hémogénique s'engagent dans la voie de différentiation hématopoïétique, ce qui est en accord avec la littérature sur le sujet [22]. Les cellules de cette sous population présentent donc des profils d'expression assez similaires les unes par rapport aux autres, une homogénéité qui les différencie par rapport aux populations d'hémangioblastes et d'endothélium hémogénique dont l'hétérogénéité a été précédemment évoquée.
Le réseau reconstitué sur les données de cette population (fig.8.7B) met en lumière la limite posée par cette homogénéité pour la reconstitution des méthodes d'apprentissage de réseaux puisqu'il n'est pas possible de déterminer les effets entre deux variables si l'une d'elle ne varie pas (ceci se démontre facilement dans le cas de l'information mutuelle, puisque $I(X;Y)$ avec $X$ constante est $I(X;Y) = -\sum_{x,y} p(x,y) log(\frac{p(x,y)}{p(x)p(y)}) = -\sum p(y) log(\frac{p(y)}{p(y)}) = 0$).
Il est également important de noter qu'étant donné le seuil assez restrictif choisi pour la reconstitution de ces réseaux (on rappelle que $Cr = 1e^{-3}$), les liens présents ici sont robustes et pourraient donc être pris en compte pour déterminer les caractéristiques du réseau de régulation de cette population. On note notamment l'importance de *Hhex*, cible de *Lmo2* et de *Etv2*, mais également activateur de plusieurs gènes, dont les facteurs de transcriptions hématopoïétiques *Sfpi1* et *Myb* (fig.8.7B).

Comme on peut le voir sur la figure 8.4C, il semble exister une différence fondamentale entre le mécanisme de différentiation des cellules de l'endothélium hémogénique et les cellules à potentiel hématopoïétique. En effet, la différentiation endothéliale apparaît asynchrone [25], remplaçant progressivement la population hémangioblastique ; tandis que la population hématopoïétique émerge rapidement et de façon synchronisée, entre les jours 7.75 et 8.25. L'importance de la diminution de l'expression des facteurs de transcription endothéliaux pour la transition endothélio-hématopoïétique (reportée dans [22]) est confirmée par notre analyse.

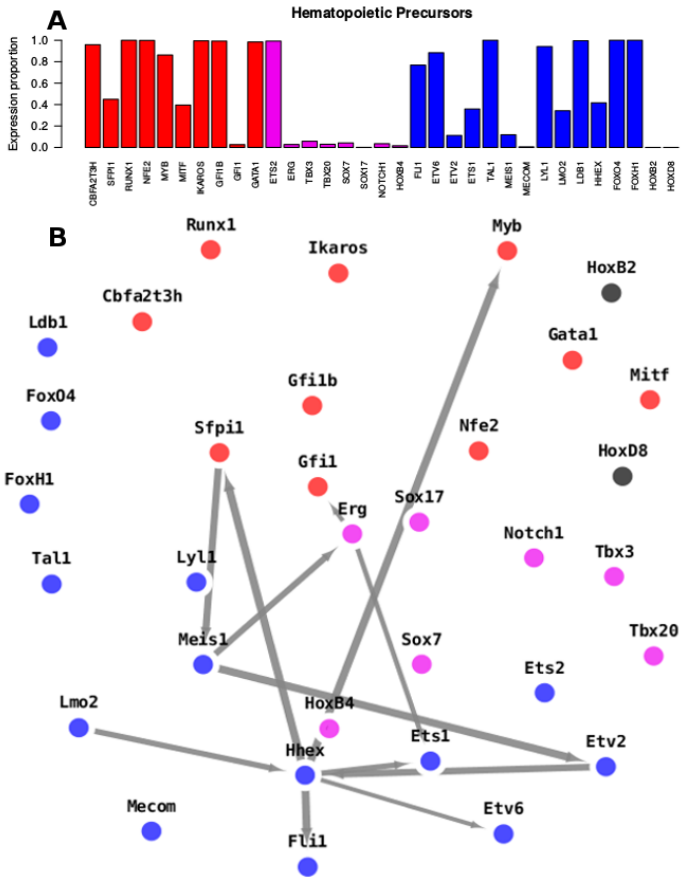



FIGURE 8.7 – **Caractérisation de la sous-population *hématopoïétique*** (**A**) Proportion de cellules exprimant chacun des 33 facteurs de transcription, répartis en 3 groupes : hématopoïétiques (en rouge), endothéliaux (en violet) et communs (en bleu). (**B**) Réseau de régulation reconstitué sur les cellules appartenant au cluster hématopoïétique, en rouge sur la figure 8.4C

## 8.4 Conclusion du chapitre

Ce travail sur l'hématopoïèse comporte donc deux partie aux objectifs distincts :

- En premier lieu, nous avons travaillé sur l'hématopoïèse chez l'adulte, un aspect relativement bien connu de ce mécanisme. La reconstitution du réseau impliquant les 18 facteurs de transcription présentés a servi à s'assurer des performances de `miic` dans un contexte de données génomiques de cellules uniques. Cette partie s'est montrée concluante puisque la capacité de `miic` à retrouver les relations transcriptionnelles connues est meilleure que la plupart des méthodes concurrentes. `miic` prédit également un certain nombre de relations transcriptionnelles qui n'ont pas été retrouvées dans la littérature, et qui constituent donc de potentielles pistes pour les équipes travaillant sur ce sujet. Ce travail a été publié dans la revue *BMC Bioinformatics* [2] (4.3).
- Dans un second temps, nous avons appliqué `miic` à un sujet beaucoup moins connu afin de d'étudier les avancées potentielles apportées par son utilisation dans ce contexte. Deux aspects se sont dégagés de ces travaux. Premièrement, sur l'ensemble des cellules, le réseau comprend plusieurs liens bi-orientés, reflétant la probable présence de variables latentes dans le système observé. De fait, l'absence d'observations concernant *Gata2*, un facteur de transcription important pour les différentiations hématopoïétique et endothéliale, et la démonstration expérimentale du rôle de celui-ci dans la co-régulation de Lyl1 et Erg [31] en font un candidat idéal pour expliquer ces liens. Deuxièmement, la reconstruction des réseaux ancestraux en complément de l'analyse en composante principale ainsi que du `k-moyens` a également permis de retrouver, mais également de nuancer le modèle de différentiation linéaire proposé dans la revue de Moignard et al., parue en 2013.

CHAPITRE 9

# PREMIER CONTEXTE CLINIQUE : PLAINTE COGNITIVE CHEZ LE SUJET ÂGÉ

---

Le déclin des capacités cognitives est fortement associé avec le vieillissement : les structures cérébrales vieillissent, et sont de moins en moins efficaces, entraînant entre autres des pertes de mémoire. Il n'est cependant pas obligatoire que cette diminution implique une perte d'autonomie significative de l'individu au quotidien. Lorsqu'on constate un affaiblissement ou une perte de plusieurs fonctions intellectuelles (cognitives) entraînant une perte d'autonomie et des troubles comportementaux, le déclin n'est plus "normal" (vieillissement physiologique), mais devient pathologique : on parle alors de syndrome **démentiel** (définition du Collège des Enseignants de Neurologie). Un certain nombre de pathologies peuvent être à l'origine de cet état, mais la majorité d'entre elles appartiennent à deux catégories distinctes :

- les démences dégénératives proviennent de la mort progressive des cellules neuronales en raison d'anomalies de la physiologie des cellules cérébrales (structures protéiques altérées, création d'amas dans les neurones, etc). Cette mort cellulaire prématurée impacte non seulement les différentes formes de mémoires, mais provoque également à terme des troubles variés tels que les chutes, les tremblements (syndrome parkinsonien) ou les hallucinations (maladie à corps de Lewy par exemple).
- les démences vasculaires ont pour origine des accidents systémiques (touchant le système vasculaire) ayant lieu dans le cerveau. Ce type de démence peut être entraînée soit par un accident vasculaire unique ayant par exemple lieu au sein d'une zone critique, soit par de multiples petits accidents qui ont pour conséquence l'apparition d'un syndrome démentiel.

En raison du vieillissement de la population (espérance de vie augmentée de 5 ans depuis

2000 selon l'Organisation Mondiale de la Santé), ces pathologies dont l'un des premiers facteurs de risque est l'âge, voient leur prévalence augmenter. Cette augmentation rend d'autant plus importante l'étude des raisons encore mal connues de leur apparition, ainsi que d'applications plus immédiates telles que l'amélioration des outils de diagnostic afin de favoriser une détection plus précoce.

C'est dans ce cadre que nous avons travaillé sur un jeu de données cliniques concernant les troubles neurologiques chez des patients âgés, afin d'explorer les avantages de l'utilisation de la reconstruction de réseaux et plus précisément de `miic`, pour la primo-analyse de données cliniques.

## 9.1 Généralités et Présentation de la base de données

Les méthodes de reconstruction de modèles graphiques ne sont communément pas utilisées sur les bases de données cliniques ; cependant ces techniques peuvent apporter une aide non négligeable pour l'appréhension de la structure des données, en mettant en évidence les dépendances directes entre les différentes variables mais également -comme on le verra dans ce chapitre- pour découvrir de nouvelles associations.
Cette partie de mon travail de thèse a donc consisté à analyser des données de consultations en neuro-gériatrie (fournies par le Pr Marc Verny, chef du service de gériatrie à la Pitié-Salpétrière) afin d'étudier l'apport de l'apprentissage de réseaux sur la visualisation, l'analyse statistique et l'interprétation des données cliniques.
Cette base de données rassemble des informations très variées recueillies sur 1628 patients ayant consulté à l'hôpital de jour de la Pitié-Salpétrière à partir de 2009 en raison de problèmes cognitifs. La collecte de ce type de données et leur rassemblement dans une base présente deux objectifs précis pour les praticiens :

- la constitution d'une banque contenant à la fois les données recueillies lors de la première consultation des patients, mais également *a posteriori*, lors de leur suivi à plus long terme lorsqu'il s'avère nécessaire (constitution de données avec séries temporelle consignant l'évolution des malades au fil du temps).
- l'utilisation de cette base comme première approche à des questions de recherche clinique ; par exemple pour bien délimiter une question et/ou définir la manière la plus pertinente de l'aborder.

A ce titre, les praticiens qui accueillent les patients participants en consultation sont chargés de recueillir différents types d'informations.

- **les antécédents du patients** : ces variables regroupent les pathologies pertinentes dont le patient a été atteint au cours de son existence. On y trouve notamment des maladies telles que les troubles anxio-dépressifs, mais également ce qui concerne les comorbidités, c'est-à-dire des pathologies chroniques n'ayant *a priori* aucun rapport avec une atteinte cognitive, mais pouvant éventuellement modifier le pronostic du patient. On peut citer parmi celles-ci le diabète, ou l'hyper tension artérielle (HTA) (fig. 9.3).
- **les résultats de l'examen clinique** : dans cette catégorie, on retrouve les conclusions du premier examen pratiqué durant la consultation, concernant par exemple la présence d'un déficit moteur ou d'un syndrome parkinsonien.
- **l'anamnèse** : c'est l'histoire des troubles du patient, mais également les résultats d'éventuels examens pratiqués auparavant, dans le cadre de la même plainte (on rappelle que les patients se présentent dans ce contexte avec une plainte concernant leur mémoire).
- **les "scores"** : ces variables comprennent l'ensemble des résultats à divers tests dont le but pour le praticien est d'avoir une idée des types de processus mnésiques atteints par la maladie du patient (mémoire de travail, mémoire épisodique ou processus attentionnels par exemple). Ces scores permettent de mieux comprendre la maladie du patient et d'affiner la recherche du diagnostic.
- **la biologie** : dans ce groupe sont rassemblées les mesures de diverses variables physiologiques relevées notamment *via* le bilan sanguin.
- **les examens d'imagerie** : ces variables indiquent les résultats et les conclusions de chacun des patients à divers types d'imagerie utilisés pour observer le cerveau. On y trouve par exemple la recherche de la trace d'accidents dans le système vasculaire du cerveau (Lacunes, AIC, microbleeds) ainsi que les résultats de tests centrés sur la recherche des marqueurs d'autres pathologies permettant d'orienter le diagnostic.
- **les traitements** : les différents médicaments que le patient prend régulièrement. Certains de ces traitements (benzodiazépines, antalgiques par exemple) peuvent avoir un impact le fonctionnement cérébral (certains processus attentionnels dans le cas des médicaments cités) ; ce qui explique l'intérêt de consigner leur prise.
- **les éléments de diagnostic** : ce sont les conclusions du médecin, après avoir "enquêté" sur l'ensemble des aspects présentés ci-dessus, au sujet de la (ou des) pathologie(s) dont est atteint le patient qui consulte. Il est tout à fait possible que

l'ensemble des résultats des différents examens et tests auxquels le patient a été soumis ne permette pas de désigner une pathologie précise. Dans ce cas le médecin effectue ce qu'on appelle un diagnostic "syndromique", c'est-à-dire qu'il restreint au mieux le champ des pathologies possibles en déterminant les fonctions mnésiques atteintes chez le patient en question. Pour cette étude, nous nous sommes concentrés sur l'étude des troubles dits "dysexécutifs", ainsi appelés en raison du fait qu'ils perturbent les fonctions du cerveau qui permettent de prendre des décisions en fonction des informations disponibles, appelées fonctions exécutives.

Nous avons donc sélectionné 91 variables représentatives de chacune de ces catégories afin de reconstruire le réseau présenté sur la figure 9.3. Nous verrons que ce type de reconstruction présente deux avantages majeurs : l'évaluation de la cohérence des données avec les connaissances sur la maladie d'une part et la découverte d'interactions peu ou non connues d'autre part.

## 9.2 Apport de la reconstitution de réseaux : Évaluation de la cohérence des données

La reconstitution du réseau lorsqu'elle est effectuée sur une quantité de données suffisante, permet la vérification à la fois des performances de la méthode (qui ont cependant déjà été explorées grâce aux différents benchmarks réalisés), mais aussi la qualité des données analysées. Dans ce but, une analyse pointue du modèle, basée sur l'exploration de la littérature ainsi que sur l'avis des médecins est réalisée. Ceci permet de s'assurer que les connaissances déjà disponibles sont effectivement restituées dans le réseau.
Dans le cas de ce jeu de données, le modèle produit par `miic` présente un ensemble de relations solides ($Cr = 1e^{-2}$, cf Partie 1), cohérentes au regard de la littérature médicale existante sur le sujet ; et de diviser le réseau en $5$ grandes sections majoritairement reliées entre elles *via* le score mesurant la dépendance du patient pour la réalisation des activités de la vie quotidienne (nœud ADL, fig. 9.3).

### 9.2.1 Variables associées aux démences provoquant un syndrome parkinsonien

Ce premier sous-ensemble regroupe les variables classiquement associées aux démences dégénératives pouvant provoquer l'apparition d'un syndrome parkinsonien ; no-

tamment la démence à corps de Lewy, ainsi que la démence parkinsonienne. Ces deux diagnostiques (LEWY et DEM_PARK, fig. 9.3) constituent la base d'une V-structure dont la pointe est la variable signalant la présence ou l'absence du syndrome parkinsonien chez le patient. Dans la littérature, le syndrome parkinsonien est défini par une rareté ainsi qu'une lenteur des mouvements, l'apparition de tremblements au repos et une rigidité ; et son apparition peut avoir plusieurs causes : une maladie de Parkinson ($80\%$ des cas), ou d'autres démences dégénératives telles que la démence à corps de Lewy ($15\%$ des cas) par exemple (données extraites de France Parkinson). La présence d'un syndrome parkinsonien peut-être détectée grâce aux résultats de l'échelle UPDRS (Unified Parkinson Disease Rating Scale), conçue dans ce but : la forte corrélation unissant ces deux variables est donc cohérente.

On retrouve, associés à la démence à corps de Lewy (DCL, variable LEWY sur la figure), les principaux symptomes de cette maladie décrits dans la littérature : Fluctuations de la cognition et de la vigilance, Troubles du Comportement du Sommeil Paradoxal (TCSP) et hallucinations [7] ainsi que l'apparition d'un syndrome parkinsonien qui constitue également l'une des conséquences de cette maladie. Les résultats du DAT-Scan (imagerie cérébrale fonctionnelle), dont l'utilisation est proposée pour le diagnostic de la DCL [28], sont aussi correctement associés à la démence à corps de Lewy sur le réseau reconstitué. Des associations indirectes (d'ordre 2) sont également intéressantes, puisqu'elles concernent certains symptomes (chutes, modifications du comportements, l'état confusionnel) sont directement associées aux principales manifestations de la DCL citées précédemment (fig. 9.3).

La démence parkinsonienne (DEM_PARK) est quant à elle liée d'une part à la présence d'une Maladie de Parkinson Idiopathique (MPI) chez le patient et d'autre part à la présence d'un syndrome parkinsonien. Ces liens qui peuvent sembler triviaux sont l'un des premiers contrôles des capacités de l'algorithme sur ce genre de données : ne pas les retrouver équivaudrait à une perte de confiance sur les autres résultats présentés. La MPI est de plus positivement associée à l'hypotension orthostatique (chute de la tension lors du passage en position debout, HTO) ce qui est également mentionné dans la littérature [38].

### 9.2.2 Variables associées à la maladie d'Alzheimer et bilan neuro-psychologique

Ce deuxième groupe est très largement composé des résultats aux tests permettant d'explorer l'état des fonctions cognitives du patient. Deux types de tests sont représentés dans ce groupe : les tests simples, constitués d'une épreuve et/ou explorant une fonction cérébrale très précise (empans directs pour les capacités attentionnelles, ou les empans indirects pour la mémoire de travail par exemple), ainsi que les tests "composites", combinant les résultats de multiples sous-tests et permettant une exploration plus globale des processus cérébraux. 3 grands tests composites sont représentés dans ce réseau : le MMS, la BREF et le MOCA. Le MMS (Mini Mental State) est un test rapide permettant de balayer de façon assez large les fonctions cognitives (fonctions liées à la mémoire, orientation temporelle et spatiale...) ; mais n'explore pas les fonctions exécutives (permettant la planification grâce au rassemblement de l'ensemble des informations disponibles). Sur le modèle reconstruit, de bons résultats au MMS sont positivement associés aux résultats de différents scores du bilan neuro-psychologique et surtout anti-corrélés via le test des 3 mots du MMS au diagnostic de la maladie d'Alzheimer. Cette partie du test MMS explore les fonctions mnésiques (c'est-à-dire liées à la mémoire), qui sont les fonctions principalement touchées dans la maladie d'Alzheimer : ce lien est donc très cohérent, de même que celui qui lie Alzheimer aux résultats des rappels (RLplusRI_diff), un ensemble d'exercice s'intéressant également aux fonctions mnésiques. Ces deux scores sont parmi les plus spécifiques de la maladie d'Alzheimer. La BREF (Batterie Rapide d'Efficience Frontale) est un test complémentaire du MMS, puisqu'il est uniquement centré sur les fonctions exécutives, localisées dans le cortex frontal : il est donc tout à fait cohérent que ses résultats soient anti-corrélés à la présence d'un trouble dysexécutif sur le réseau(fig.9.3). Le MOCA (Montreal Cognitive Assessment) est un score compilant en réalité les résultats du patients pour différents tests, notamment le test dit de l'horloge (le patient doit dessiner une horloge analogique indiquant 11h10) ou un test de fluence sémantique (le patient doit donner un maximum de mot commençant par la lettre "F" en une minute), expliquant les liens reliant le MOCA avec ces tests.

### 9.2.3 Variables associées à l'état psychiatrique du patient

Dans ce groupe sont rassemblées toutes les variables liées à l'état psychiatrique. Il est composé des maladies psychiatriques, présentes ou passées (antécédents psychiatriques,

anxio-dépression et syndrome bipolaire), des traitements associés à ces maladies (anti dépresseurs, psychotropes, benzodiazépine, neuroleptiques) et enfin des scores permettant de mettre en évidence une dépression (GDS_15) et une dégradation de la qualité de vie (Qol).
L'analyse de l'ensemble des liens reliant ces variables permet de confirmer la cohérence de l'ensemble : une bonne qualité de vie est étroitement associée à un faible score GDS 15 (correspondant à une faible probabilité de dépression). Les pathologies psychiatriques sont également toutes reliées les unes aux autres,

### 9.2.4 Variables associées aux démences vasculaires

Ce quatrième regroupement de variable associe les variables jouant un rôle dans un autre type de démence, appelé démence vasculaire. Ces démences sont provoquées par une pathologie vasculaire au niveau cérébral : l'accident endommage des zones du cerveau indispensable aux processus cognitifs, provoquant une démence (d'origine vasculaire donc) chez le patient.
Ce groupe contient donc tous les types d'accidents vasculaires (distingués en fonction de leur taille et de leur type) : l'accident vasculaire cérébral (AVC), l'accident ischémique cérébral (AIC), les micro-saignements cérébraux (microbleeds) et les lacunes. La succession de ce type de pathologie peut provoquer des démences vasculaires, qui peuvent également être associées à une démence dégénérative ; on parle alors de démence mixte (FORM_MIXTES).

### 9.2.5 Variables associées aux comorbidités

Ce dernier regroupement comprend les variables associées aux comorbidités, c'est-à-dire des maladies différentes (hypertension artérielle (HTA), diabète, bronchopneumopathie chronique obstructive (BPCO)), mais qui sont susceptibles d'avoir un impact sur le pronostic vital du patient. L'ensemble des liens reliant ces variables est très cohérent, chaque pathologie étant associée à son traitement (la variable diabète est associée à la prise d'Anti-Diabétiques Oraux par exemple) et/ou à l'examen permettant de la mettre en évidence (Arythmie Cardiaque par Fibrillation Auriculaire (ACFA) est associée aux résultats de l'électrocardiogramme (ECG)).

### 9.2.6 Liens bi-orientés

On retrouve également dans ce réseau 4 liens doublement orientés, trace possibles de la présence d'une variable non observée influençant les deux nœuds reliés. La plus notable d'entre elles concernent les variables "Tabac_actif" et "Tabac_sevré". L'explication est ici logique : la cause commune de ces deux variables pourrait très vraisemblablement être nommée "Tabac", rassemblant ainsi les patients ayant fumé pendant une période (terminée ou non) de leur vie.
Un autre lien doublement orienté relie les variables "AAP" (anti-agrégants plaquettaires) et "AVK" (anti-vitamine K). Ces deux traitements tentent à limiter la coagulation du sang chez les patients sujets aux accidents vasculaires, et ne sont en général pas associés chez un même patient. La variable latente ici pourrait être la décision médicale de prescrire l'un ou l'autre de ces deux types de traitements.

## 9.3 Apport de la reconstitution de réseaux : Découverte de relations non-triviales

En médecine et en étude clinique, pour découvrir une association nouvelle entre deux variable d'intérêt, l'usage est d'abord de formuler l'hypothèse, puis de vérifier cette hypothèse spécifique sur un jeu de données, obtenu soit par l'intermédiaire d'une cohorte pré-existante, soit en élaborant une étude clinique basée sur cette hypothèse.
Le second apport de l'analyse de ce type de données grâce aux méthodes de reconstruction de réseaux est la mise en évidence de liens non triviaux, sans avoir besoin nécessairement de connaissance *a priori*. En effet, la reconstitution d'un réseau s'apparente à une analyse multivariée effectuée sur toutes les variables du jeu de données à la fois ! Ainsi, certaines des associations retrouvées dans notre modèle n'ont été publiées que récemment ; c'est le cas de l'association mise en évidence entre les résultats des examens d'imagerie des échelles de Fazekas et de Scheltens (fig. 9.3, anneau vert).
L'échelle de Fazekas, proposée par Fazekas et al en 1987 [9], vise à quantifier l'importance de brillances blanches appelés hypersignaux (fig.9.1), observées grâce à l'Imagerie à Résonnance Magnétique (IRM) de la substance blanche du cerveau des patients, en leur attribuant un score reflétant leur sévérité. Les hypersignaux de la matière blanche, souvent observés sur les IRM du cerveau des patients âgés, sont le reflet d'une démyélinisation (perte de l'enveloppe protectrice des cellules neuronales), ainsi que d'une perte axonale

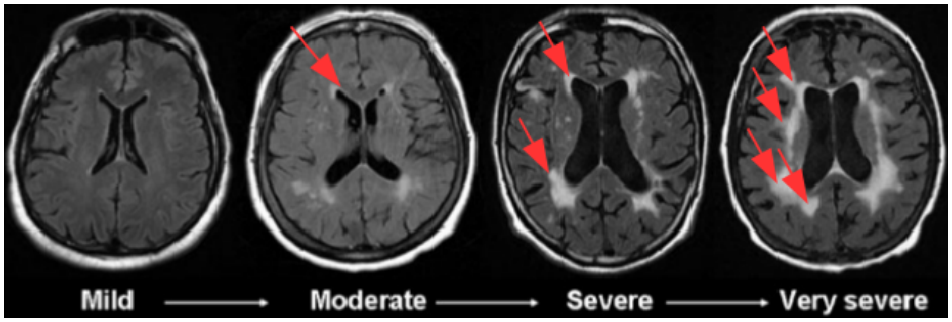



FIGURE 9.1 – **Echelle de sévérité d'anomalies de la substance blanche** – adapté de Chutinet et Rost,*Curr Treat Options Cardio Med*, 2014. Les flèches rouges pointent des hyperintensités de la substance blanche

causées par de petites pathologies vasculaires cérébrales. La prévalence et la sévérité de ces lésions augmente avec l'âge, mais également avec l'hypertension artérielle et leur présence accompagnée d'autres symptômes tels que les lacunes ou les micro-saignements cérébraux est indiquée comme les premiers signes de l'apparition d'une démence vasculaire[32]. L'échelle de Fazekas aide donc à déterminer si leur nombre ainsi que leur importance est uniquement la conséquence du vieillissement physiologique du cerveau ou si une (ou des) pathologie(s) doivent être envisagées pour expliquer leur sévérité.
L'échelle de Scheltens, proposée par Scheltens et al. en 1992 ([36]), vise quant à elle à attribuer un score en fonction de la sévérité d'une éventuelle atrophie de l'hippocampe (9.2), une structure du cerveau impliquée dans la mémoire et la navigation spatiale particulièrement touchée dans de multiples pathologies, dont la maladie d'Alzheimer ([12]). La mesure de son volume permet d'aider au diagnostic : l'atrophie de l'hippocampe est en effet un signe recherché dans le cas de pathologies diverses telles que la dépression ou la maladie d'Alzheimer par exemple ([39]).

On remarque tout d'abord que les résultats à l'échelle de Fazekas sont les conséquences directe des variables démence vasculaire (DEM_VASC) et formes de démences mixtes (FORM_MIXTES), ce qui est cohérent avec la littérature[32]. On remarque également une association significative entre les résultats des deux échelles, association qui n'était pas publiée dans le contexte de la démence chez le sujet âgé avant mars 2017[11]. On note toutefois que dans le but de mettre en évidence de telles interactions, les méthodes classiquement employées nécessitent la formulation d'une question précise explorée par diverses analyses statistiques. La reconstruction de réseaux ancestraux telle que proposée par `miic` permet donc de contourner ce genre de difficulté en explorant l'ensemble des

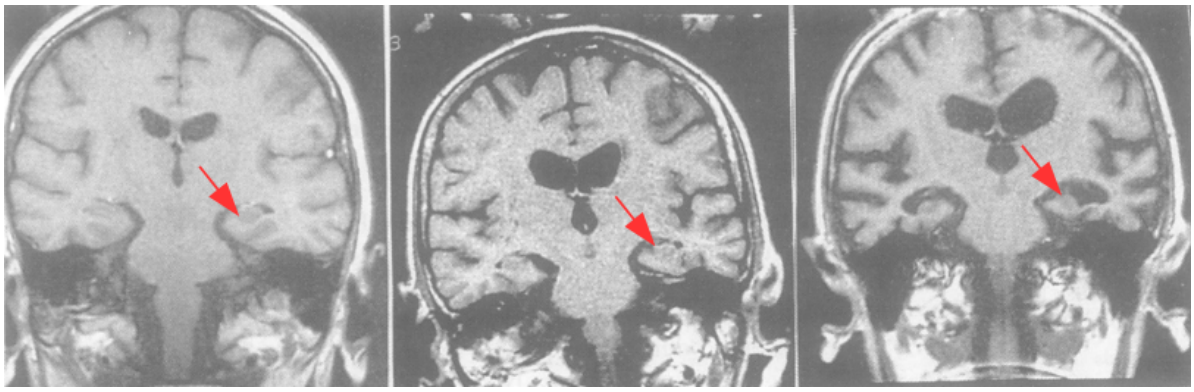


FIGURE 9.2 – **Exemple montrant une atrophie croissante (de gauche à droite) de l'hippocampe** – adapté de Scheltens et al.„*Journal of Neurology, Neurosurgery, and Psychiatry*, 1992. Les flèches rouges pointent les hippocampes dont le volume décroît de manière visible.

interactions directes présentes dans un jeu de données.

## 9.4 Conclusion du chapitre

Deux bénéfices principaux se dégagent donc de l'application de la reconstruction de réseaux ancestraux dans un contexte de données cliniques. Premièrement, la robustesse de la méthode ainsi que la cohérence du réseau observé permet d'effectuer un premier contrôle qualité sur la base de données utilisée : si un lien est absent, on peut aisément comprendre pourquoi en analysant les informations détaillées de la reconstruction du réseau, mais également en observant directement le jeu de données. Deuxièmement, la mise en évidence, sans nécessité de connaissances ni d'hypothèses *a priori*, d'associations ignorées jusqu'à maintenant pourrait également constituer un avantage considérable dans la primo-analyse de ce genre d'ensemble de données variées et complexes.

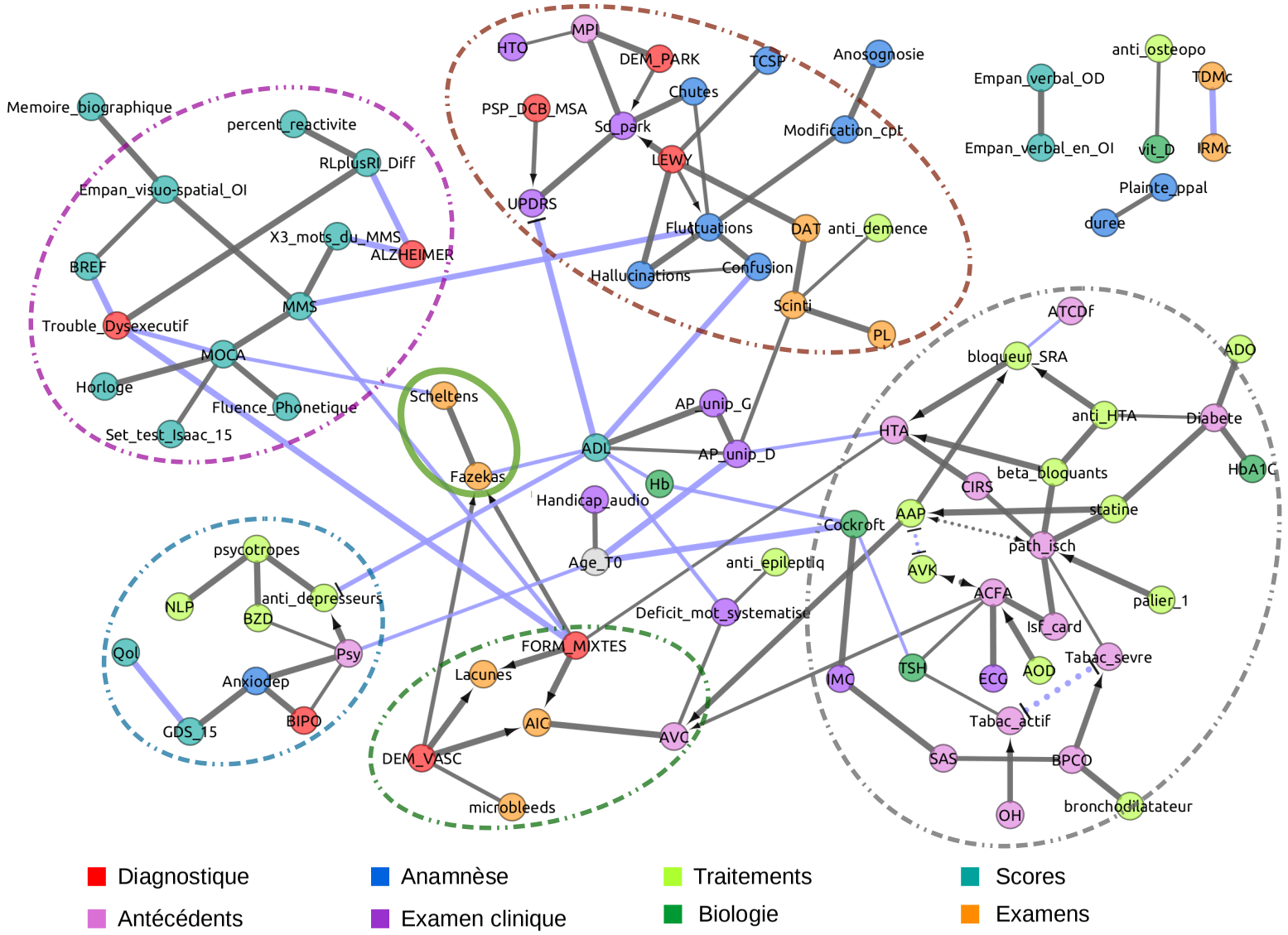



FIGURE 9.3 – **Réseau de relations entre les caractéristiques des patients âgés consultant pour des problèmes cognitifs.** Orange : Variables liées aux démences dégénératives autres que la maladie d'Alzheimer. Violet : Composantes corrélées (directement ou indirectement) à la maladie d'Alzheimer. Bleu ciel : Variables associées à l'état psychologique du patient Vert : Variables associées aux démences vasculaires. Gris : Variables décrivant les comorbidités, ainsi que les traitements associés. Anneau Vert : Association entre les résultats aux échelles de Fazekas et de Scheltens

CHAPITRE 10

# SECOND CONTEXTE CLINIQUE : CANCER DU SEIN

## 10.1 Généralités

Le cancer du sein est l'un des plus fréquent chez la femme et même si la mortalité tend à diminuer, sa fréquence n'en fait pas moins un enjeux majeur de santé publique. En collaboration avec l'équipe du Residual Tumor and Response to Treatment Laboratory (Dr. Fabien Reyal et Dr. Anne-Sophie Hamy-Petit), nous avons pu appliquer l'algorithme `MIIC` sur les données issues de la cohorte *Neorep*, qui rassemble des patientes ayant été atteintes d'un cancer du sein. Nous avons décidé de suivre les associations entre divers types de variables susceptibles de jouer un rôle clé dans le pronostic de la maladie :

- **Informations générales sur la patiente** : Ces nœuds correspondent à des variables décrivant la patiente ($Age$, $Menopause$,Body Mass Index ( $BMI$)...), mais également des variables très globales sur la tumeur contractée, telle que la taille de celle-ci ($cT$) ou **l'envahissement ganglionnaire** ($cN$) qui indique si la tumeur a atteint ou non les **ganglions**.
- **Niveau de vie** : Sous l'impulsion du Dr Reyal, nous avons également intégré de nouvelles informations sur le niveau de vie des patientes à la base de données, afin de voir si celles-ci avait un impact quelconque sur les conséquences de la maladie par exemple. Nous avons ainsi pu ajouter les variables $DistanceAuCentre$, qui rend compte de la distance entre le domicile de la patiente et l'endroit où elle est suivie ainsi que $RevenuMed$, le revenu médian dans la commune d'origine de la patiente.
- **Caractéristiques histologiques de la tumeur** : Ce groupe rassemble les variables décrivant les caractéristiques "visuelles" de la tumeur, à savoir si les cellules ont tendances à se diviser très rapidement ($Mitotic\_index$) ou si l'on retrouve des cellules cancéreuses dans les vaisseaux avoisinants ($embols\_biopsie$, $embols\_sein$).

La variable $Grade$ rend compte du niveau d'agressivité de la tumeur (3 grades, par ordre d'agressivité croissante).

- **Immunologie** : Les nœuds appartenant à cette catégorie caractérisent l'évolution des cellules immunitaires (TILs, *Tumor Infiltrating Lymphocytes*) présentes au sein de la tumeur.
- **Variables prédictives et traitements associés** : deux traitements sont évoqués sur ce réseau :
  - **Hormonothérapie** : regroupement des variables associées au traitement par hormonothérapie, c'est-à-dire le traitement lui même ainsi que le statut hormonal de la tumeur, $ER$ et $PR$, indiquant la présence sur les cellules tumorales de récepteurs à œstrogènes et à la progestérone, respectivement.
  - **Herceptine** : regroupe les deux variables associées au traitement par le trastuzumab (également appelé Herceptine), c'est-à-dire là encore le traitement lui-même ($trastuzumab$), ainsi que le statut du gène HER2, conditionnant l'efficacité du traitement.
- **Issue de la maladie** : Ce groupe rassemble la caractérisation de la réponse au traitement ($pCR$), les différents délais avant une rechute, l'apparition de métastases ($Delai\_PremRecidive$, $Delai\_Meta$) ainsi que l'information sur le décès éventuel de la patiente ($Death$).

L'ensemble de ces variables a donc été utilisé pour s'assurer de la cohérence du modèle graphique inféré, mais également pour explorer de nouvelles pistes telles que l'impact du niveau de vie de la patiente sur la maladie par exemple.

## 10.2 Apport de la visualisation des données via un modèle graphique

De la même manière que dans le chapitre 9, nous allons montrer dans cette section que la reconstruction de réseaux sur les données de *Neorep* a permis d'une part d'effectuer un premier contrôle de la qualité de ces données (absence de relations connues dues à un biais de sélection par exemple) et d'autre part de rechercher les associations connues (publiées ou en passe de l'être). Les interactions figurant sur le réseau reconstruit (fig 10.1) présentent ces deux intérêts, puisqu'on retrouve des associations évidentes, confirmant la robustesse des données ainsi que de la méthode et des interactions récemment décou-

vertes *via* les techniques d'analyse classiques appliquées précédemment par l'équipe du Dr Reyal.

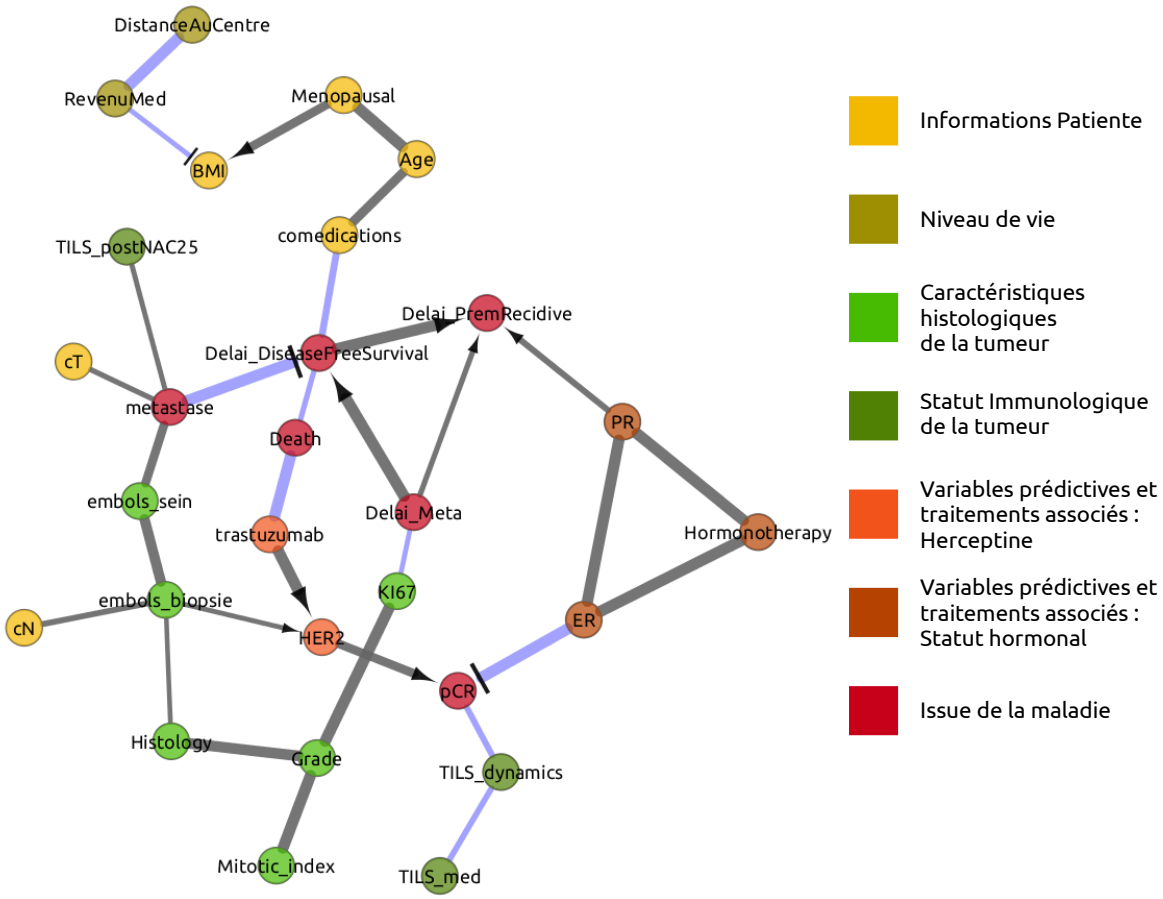



FIGURE 10.1 – **Réseau reconstruit à partir des données de la base *Neorep*** – Le réseau est filtré avec $C_r = 10^{-1}$. Les associations colorées en bleu correspondent à des anti-corrélations, tandis que les grises soulignent des corrélations positives.

### 10.2.1 Liens classiques, contrôle qualité de la base *Neorep*

La présence des interactions détaillées dans cette section montrent que le jeu de données est cohérent avec les connaissances publiées sur le sujet.

- **Association** $Age - Menopause$ : L'association positive entre ces deux variables est une évidence et c'est un marqueur de qualité qu'il est important de retrouver

pour souligner le fonctionnement correct de la méthode `MIIC`.

- **Association** $Age - Comedication$ : Le nœud $Comedications$ représente le fait qu'une patiente suit (ou non) un traitement, en dehors de celui administré pour son cancer. Il est évident que plus l'âge augmente, plus les patientes sont susceptibles de prendre des traitements médicamenteux, ce qui explique cette association.
- **Association négative** $RevenuMed \rightarrow BMI$ : La variable $RevenuMed$ indique la tranche de revenue médiane dans laquelle est située la commune d'origine de la patiente (indicateur du niveau de vie), tandis que $BMI$ est l'abréviation de **Body Mass Index**, un indicateur de l'obésité. Le réseau indique qu'un faible revenu semble favoriser un fort BMI, ce qui a déjà été observé dans d'autres études (Food Research and Action Center). Enfin, on note que la v-structure $RevenuMed \rightarrow BMI \leftarrow Menopausal$ confirme l'indépendance attendue entre les variables $RevenuMed$ et $Menopausal$.
- **Association** $Ménopause \rightarrow BMI$ : L'association entre ces deux variables laisse penser que la ménopause favorise (directement ou à travers un intermédiaire caché) un BMI plus important. Cette association mentionnée dans la littérature, même si les auteurs de cette étude la qualifie de ténue ([24]). La **v-structure** ayant pour pointe la variable $BMI$ est globalement cohérente, puisqu'elle indique une **indépendance marginale** entre les revenus des patientes et le fait qu'elles soient ménopausées.
- **Triangle** $PR - ER - Hormonotherapy$ : Les sigles $PR$ et $ER$ signifient **Progesteron Receptor** et **Estrogen Receptor**. Ces variables indiquent la présence de récepteurs à la progestérone et aux œstrogènes, deux hormones, à la surface des cellules tumorales. Une tumeur exprimant ce genre de récepteurs (en particulier les ER), est appelée **hormonosensible**, c'est-à-dire que son développement est favorisé par ces hormones, naturellement produites par le corps humain. Il existe donc un traitement, appelé **hormonothérapie**, visant à éviter cet effet stimulant sur la tumeur, soit en diminuant la quantité d'hormones produites par le corps, soit en bloquant directement leur action sur la tumeur.
- **Groupe** $Histology - Grade - Ki67 - Mitotic\_index$ : Le grade d'une tumeur est une classification de l'agressivité de celle-ci. Il est défini en prenant en compte l'apparence des cellules ainsi que leur activité mitotique. Les trois interactions retrouvées sur le réseau, centrées sur $Grade$ reflètent ce processus de définition : l'intensité de l'activité de division cellulaire (caractérisée par les variables $Ki67$ et

$Mitotic\_index$) et l'aspect des cellules au microscope ($Histology$) sont associés au grade de la tumeur, qui permet de rendre compte de toutes les corrélations entre les variables qui lui sont associées.

### 10.2.2 Interactions publiées sur la base *Neorep*, ou dans des conditions similaires

- $PR - Delai\_PremRecidive$ : Le réseau montre que lorsque la tumeur présente des récepteurs à la progestérone, le temps entre la fin du traitement et la première rechute (si elle a lieu) a tendance a être plus important. Cet effet a déjà été découvert et publié par Baulies et al. en 2015 ([35]).
- $Delai\_Meta - Delai\_PremRecidive$ : L'interaction causale retrouvée entre le délai sans métastases (cellules cancéreuses ayant quitté le site de la tumeur initiale, et pouvant donc "propager" le cancer à d'autres organes), et le délai sans rechute a également été découverte par Baulie et al en 2015 ([35]) et confirmée grâce à une nouvelle analyse effectuée sur *Neorep* dans un article à paraître. Il est cependant important de noter que la causalité inférée sur le réseau ne semble pas en accord avec les publications, puisqu'elles indiquent que c'est plutôt une récidive survenue rapidement qui peut favoriser l'apparition de métastases et non l'inverse, comme figuré sur le graphe. Une erreur d'orientation n'est donc pas à exclure dans ce cas.
- $Delai\_DiseaseFreeSurvival - Death - Trastuzumab \rightarrow HER2 \rightarrow pCR$ : Ce motif caractérise une association entre l'administration du trastuzumab (aussi appelé herceptine) et la *disease free survival* (survie sans maladie, DFS) à travers la variable $Death$, rendant compte de l'éventuel décès de la patiente. Le réseau montre donc que les patientes à qui l'on administre cette molécule sont moins susceptibles de mourir. On trouve également une association positive indirecte entre les variables $Trastuzumab$ et $pCR$ (pathological Complete Response, indiquant une réponse favorable au traitement). Cette association est médiée par la variable $HER2$, qui indique la surexpression du gène HER2 dans la tumeur. Bien que cette caractéristique favorise la prolifération cellulaire et donc le développement du cancer, elle rend également le traitement par trastuzumab beaucoup plus efficace. Le sens de la causalité entre les variables $Trastuzumab$ et $HER2$ est cependant *a priori* de nouveau erroné, puisqu'il semble que ce soit la surexpression du gène HER2 qui entraîne l'administration d'un traitement par trastuzumab, et non l'in-

verse [15].

- $TILS\_postNAC25 - Metastase - Delai\_DiseaseFreeSurvival$ : La variable $TILS\_postNAC25$ indique si des Tumor Infiltrating Lymphocytes (cellules immunitaires ayant infiltré la tumeur, TILs) sont observés après la chimiothérapie néo-adjuvante (NAC). Le réseau montre une association entre cette variable et la variable $metastase$ (indique si la tumeur de la patiente a ou non métastasé), qui est elle même associée négativement avec la variable DFS précédemment évoquée. Une interaction $TILS\_postNAC25 - Delai\_DiseaseFreeSurvival$, vient d'être publiée par Hamy et al [14], à partir d'analyse multivariée sur la base *Neorep*. Cette même étude mentionne une association $TILS\_postNAC25 - metastase$ significative ; ce qui est en accord avec le réseau reconstitué. Il est cependant intéressant de noter que l'interaction entre la présence de TILs et la DFS est médiée par la variable $metastase$ sur le réseau. Cette observation pourrait ouvrir de nouvelles pistes quant à l'interprétation de la relation découverte par Hamy et al [14].
- $pCR - TILS\_dynamics$ : Hamy et al mentionnent également dans leur étude sur la base *Neorep* une interaction entre la $pCR$ et $TILS\_dynamics$ (qui correspond à la différence entre le nombre de TILs observés avant et après la chimiothérapie néoadjuvante)(10.2.2). Le réseau rend exactement compte de cette interaction au travers d'une association **négative** entre ces deux variables (fig.10.1).


FIGURE 10.2 – **Extraite de Hamy et al[14]** – Association entre le changement du nombre de TILs (représenté en pourcentage de différence par rapport au nombre pré traitement) et la réponse à la chimiothérapie néoadjuvante.

- $cN - embols\_biopsie - HER2$ : Le nœud $cN$ représente **l'envahissement ganglionnaire**, c'est-à-dire qu'il indique si la tumeur a ou non atteint des structures appelées ganglions lymphatiques ; tandis que la variable $embols\_biopsie$ indique si des cellules tumorales ont été observées dans les vaisseaux sanguins et/ou lymphatiques situés en périphérie de la tumeur. Les associations entre l'observation de ces emboles et l'envahissement ganglionnaire, ainsi que la surexpression du gène HER2 dans la base *Neorep* ont également été mises en évidence par l'équipe de recherche de l'hôpital de l'institut Curie qui réalise les analyses sur cette base (thèse du Dr Prescilla PAIS).

### 10.2.3 Interaction négative entre les comédications et la DFS : Hypothèse d'une variable cachée

Une des association montrée par le graphe n'a pas encore été publiée, et peut éventuellement devenir une piste pour des recherches futures sur le sujet ; c'est l'interaction négative entre les variables $comedications - Delai\_DiseaseFreeSurvival$. La *disease free survival* (survie sans maladie, DFS) est la période pendant laquelle aucun symptôme du cancer n'est observé, après la fin d'un traitement. Cette association signifie donc que les patientes de la base *Neorep* prenant des traitements non associés au cancer ont généralement une DFS plus courte. L'une des explications possibles de cette interaction est que l'état de santé général d'une patiente influe sur la DFS. Cette hypothèse pourrait être vérifiée en introduisant une variable de type $EtatDeSanteGeneral$ qui, si cette explication s'avère correcte, devrait être un intermédiaire, entraînant l'apparition du motif suivant : $comedications - EtatDeSanteGeneral - Delai\_DiseaseFreeSurvival$.

## 10.3 Conclusion de l'analyse par MIIC de la base de données*Neorep*

Cette application montre clairement l'avantage que peut avoir une analyse en reconstruction de réseaux dans le contexte de données cliniques. En effet, sans effectuer la moindre hypothèse *a priori* et sans avoir à isoler de problématique particulière, le réseau reconstitué met en évidence un grand nombre des relations découvertes au fil du temps par les chercheurs de l'hôpital de l'Institut Curie. Ceci souligne également la robustesse

de l'analyse effectuée, puisque le réseau était filtré ($C_r = 10^{-1}$), et que la base de données contient un très grand nombre de valeurs manquantes.

CHAPITRE 11

# CONCLUSION

Ce manuscrit présente l'algorithme `MIIC`, une nouvelle méthode d'apprentissage de modèles graphiques avec variables latentes. La méthode `3off2`, sur laquelle repose `MIIC`, propose une reconstruction de modèles graphiques en suivant le schéma en trois étapes des méthodes basées sur la contrainte :

1. Inférence du **squelette** du réseau grâce à la découverte progressive des indépendances conditionnelles entre les variables.
2. Orientation des motifs appelés v-structures, seules traces de la causalité dans des données d'observations.
3. Propagation de ces orientations, en suivant les règles d'orientation des CPDAG $R_{1:3}$, permettant d'éviter la création de nouvelles v-structures et de cycles orientés.

Lors de l'inférence du squelette, la sélection des ensembles de séparation réalisée par `3off2` pour la découverte des indépendances conditionnelles grâce à un score ($R$, basé sur la théorie de l'information) permet d'éviter l'écueil algorithmique représenté par la recherche aléatoire des ensembles de séparation, effectuée par les méthodes basées sur la contrainte classiques.

`MIIC` comporte plusieurs améliorations par rapport à `3off2`, dont les plus notables sont :

$i$ la possibilité de rechercher des causes latentes,

$ii$ la possibilité de filtrer les liens inférés afin de ne retenir que ceux possédant la plus faible probabilité d'être retenus par hasard

$iii$ la possibilité de corriger l'apprentissage en prenant en compte le nombre d'échantillons véritablement indépendants dans le jeu de données

Concernant la prise en compte des variables latentes et donc la possibilité de reconstruire des PAGS, nous avons réalisé des études comparatives sur trois réseaux benchmarks de difficultés variées : Alarm (37 nœuds, 46 liens, 509 paramètres), Insurance (27 nœuds, 52 liens, 984 paramètres) et Barley (48 nœuds, 84 liens, 114005 paramètres). Les résultats

de ces études montrent que les performances de MIIC sont meilleures à la fois en terme de résultats (Précision, Recall, Fscore), mais également en terme de temps de calculs 7.2.

Sur le plan pratique, ce manuscrit démontre l'utilité de la reconstruction de réseaux causaux en prenant en compte l'effet des variables latentes à travers trois applications distinctes. En premier lieu, l'inférence du modèle de régulation transcriptionnelle de l'hématopoïèse adulte nous a permis de démontrer l'efficacité et la robustesse de notre méthode sur des données réelles, avec l'ensemble des biais que cela peut comporter. La deuxième partie de cette application sur l'hématopoïèse a permis d'explorer un jeu de données regroupant des cellules à différents stades de leur différentiation en cellules hématopoïétiques primitives. La reconstruction du réseau constitué de 33 facteurs de transcription sur l'ensemble des 3934 cellules extraites par Moignard et al. nous a permis de souligner l'importance de la prise en compte de l'effet des variables latentes, au travers de liens bi-dirigés robustes et pouvant être expliqués par l'absence d'un facteur de transcription majeur, *Gata2*. De plus, l'utilisation de méthodes d'analyses complémentaires telles que la PCA ou le clustering nous ont permis de retrouver le schéma général de différentiation de ces cellules hématopoïétiques, tout en y ajoutant un certain nombre de nuances, notamment au sujet de l'importance de l'inhibition des gènes endothéliaux dans ce processus.
Dans un second temps, nous avons pu appliquer MIIC a des jeux de données cliniques, grâce à deux collaborations, la première avec le service de gériatrie de la Pitié-Salpétrière (Pr. Marc Verny) et la seconde avec l'équipe du Residual Tumor and Response to Treatment Laboratory (Dr. Fabien Reyal et Dr. Anne-Sophie Hamy-Petit).

Le jeu de données concernant les plaintes cognitives chez le sujet âgé a permis la reconstruction d'un réseau de 91 variables, couvrant à la fois l'état général du malade, mais également le diagnostic de diverses pathologies neurologiques ou autres (comorbidités) ainsi que leurs symptômes et le traitement éventuellement associé. Ce modèle est très cohérent avec la littérature, autant sur le plan du squelette que sur le plan des orientations. Il présente également des interactions très récemment découvertes par les spécialistes du domaine, telle que l'association entre les hypersignaux dans la matière blanche et l'atrophie de l'hippocampe, publiée en mars 2017. L'importance de la prise en compte des effets des variables latentes y est également soulignée, grâce à l'inférence de plusieurs liens bi-dirigés tel que celui entre les variables $tabac\_actif$ et $tabac\_sevré$ par exemple, pour lesquelles le principe de la cause commune est relativement simple à démontrer.

La seconde application clinique porte sur le cancer du sein, grâce à l'analyse des données de la cohorte *Neorep* qui rassemble des patientes atteintes de cette pathologie dans le but d'étudier l'impact d'un traitement, l'herceptine, en chimiothérapie néo-adjuvante. Le réseau de 27 variables inféré sur ces données aide également à souligner l'utilité de `MIIC` pour le contrôle qualité des données, grâce à l'inférence de liens évidents mais qu'il est indispensable de retrouver notamment. Le second aspect intéressant de cette reconstruction réside dans le fait que sans hypothèse *a priori*, la méthode a été capable de retrouver des associations complexes publiées très récemment par l'équipe du Residual Tumor and Response to Treatment Laboratory, en charge de l'analyse de cette cohorte.

Sur le plan des perspectives, plusieurs publications sont en préparation. La première concerne la création d'un serveur permettant d'utiliser `MIIC` en ligne. La seconde publication en préparation porte sur l'élargissement de cette méthode aux données continues. En effet, l'information mutuelle sous la forme présentée dans ce manuscrit, n'est calculée que sur des variables discrètes ; mais l'équipe a terminé la mise au point d'une technique permettant de découvrir une discrétisation **optimale**, maximisant l'information mutuelle du lien testé. Enfin, une application sur la possibilité de prédire les contacts entre les résidus d'une séquence protéique donnée est étudiée.

QUATRIÈME PARTIE

# Annexes

## 11.1 Lexique : sigles des variables du réseau présenté dans le chapitre 9

### 11.1.1 Groupe Syndromes Parkinsoniens

- **HTO** : Hypotension orthostatique
- **MPI** : Maladie de Parkinson Idiopathique
- **DEM_PARK** : Démence parkinsonienne
- **Sd_park** : Syndrome parkinsonien
- **TCSP** : Troubles Comportementaux du Sommeil Paradoxal
- **DAT** : Dat-Scan
- **PL** : Ponction Lombaire
- **UPDRS** : Unified Parkinson Disease Rating Scale
- **PSP** : Progressive Supranuclear Palsy
- **DCB** : Dégénérescence Cortico-Basale
- **MSA** : Multiple System Atrophy
- **Scinti** : Scintigraphie

### 11.1.2 Groupe Maladie D'Alzheimer

- **BREF** : Batterie Rapide d'Efficience Frontale
- **MOCA** : Montreal Cognitive Assessment
- **MMS** : Mini-Mental State

### 11.1.3 Groupe Psychiatry

- **Qol** : Quality of life
- **NLP** : Neuroleptiques
- **BZD** : Benzodiazépine
- **Anxiodep** : Anxiodépressif

### 11.1.4 Groupe Vascular

- **AVC** : Accident Vasculaire Cérébral
- **AIC** : Accident Ischémique Cérébral

- **DEM_VASC** : Démence vasculaire
- **FORM_MIXTES** : Démences mixtes

### 11.1.5 Groupe Comorbidity

- **IMC** : Indice de Masse Corporelle
- **SAS** : Syndrome d'Apnée du Sommeil
- **OH** : Consommation d'alcool
- **BPCO** : Bronchopneumopathie chronique obstructive
- **ECG** : Électrocardiogramme
- **ACFA** : Arythmie Cardiaque par Fibrillation Auriculaire
- **AOD** : Anticoagulant Oral Direct
- **AVK** : Anti-Vitamine K
- **AAP** : Anti-agrégant Plaquettaire
- **HTA** : Hypertension artérielle
- **CIRS** : Cumulative Illness Rating Scale
- **ADO** : Antidiabétiques oraux
- **HbA1c** : Hémoglobine glyquée
- **bloqueur_SRA** : Bloqueurs du Système Rénine-angiotensine
- **anti_HTA** : Anti Hypertenseurs
- **ATCDf** : Antécédents Familiaux

### 11.1.6 Autres

- **vit_D** : Vitamine D
- **TDMc** : Tomodensitométrie Cérébrale
- **IRMc** : Imagerie par Résonnance Magnétique Cérébrale
- **Hb** : Hémoglobine
- **AP_unip_D** : Appui Unipodal Droit
- **AP_unip_G** : Appui Unipodal Gauche
- **ADL** : Activity of Daily Living

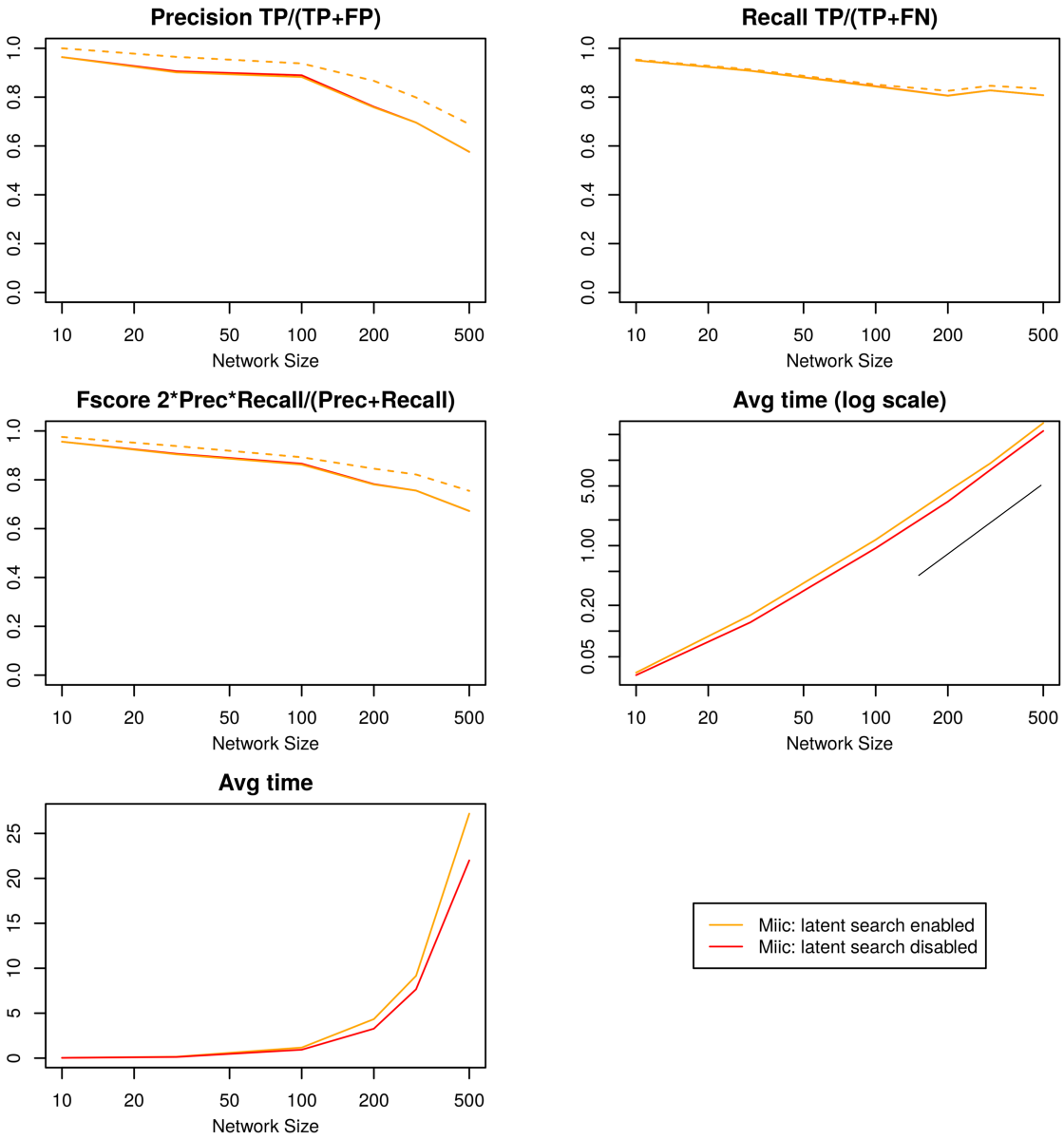



FIGURE 11.1 – **Temps de calcul en fonction du nombre de variables** – avec ou sans recherche de variables latentes. Les Fscores sont identiques, puisque ces tests ont été effectués sur des DAG. Les échelles de temps sont très similaires, avec une évolution quadratique en échelle logarithmique en fonction du nombre de nœuds.

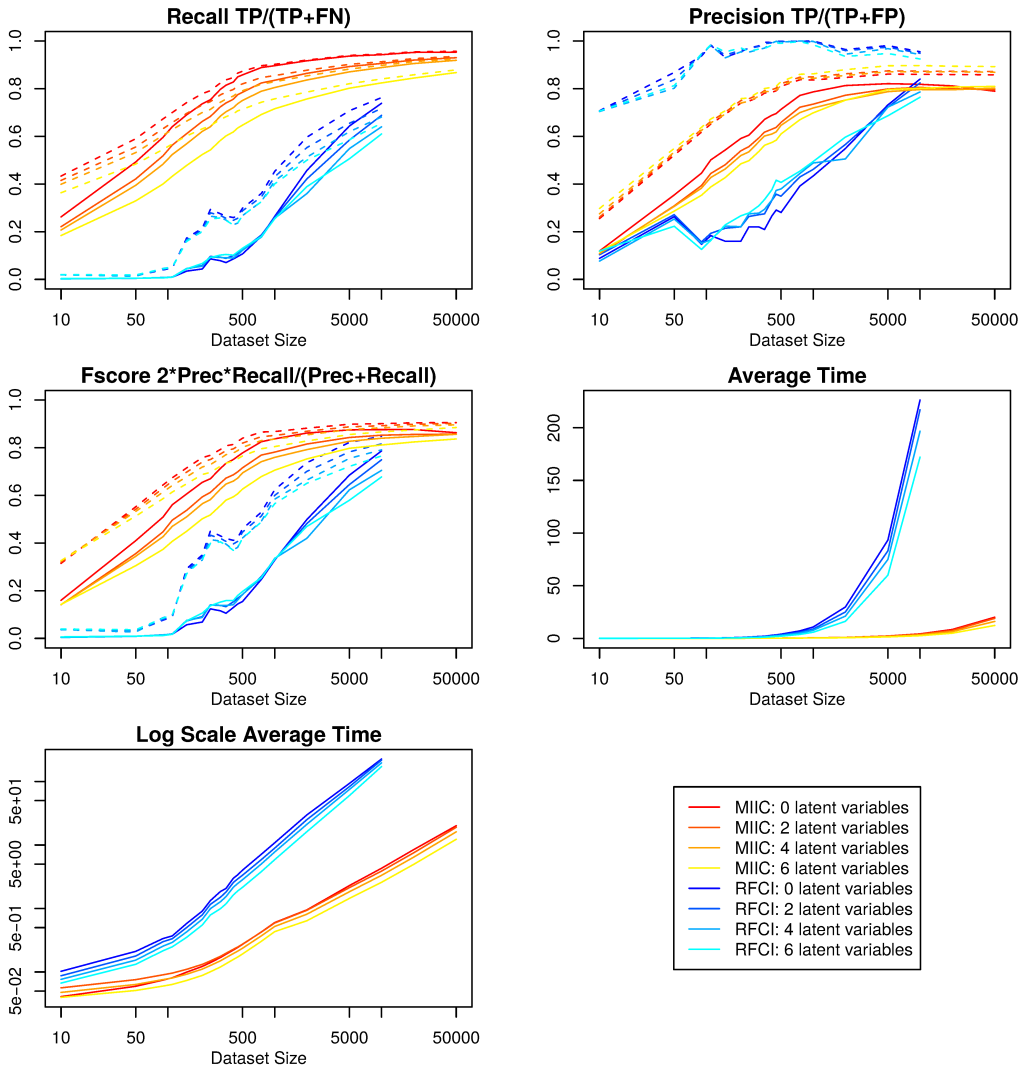



FIGURE 11.2 – **Résultats complets : réseau benchmark Alarm**

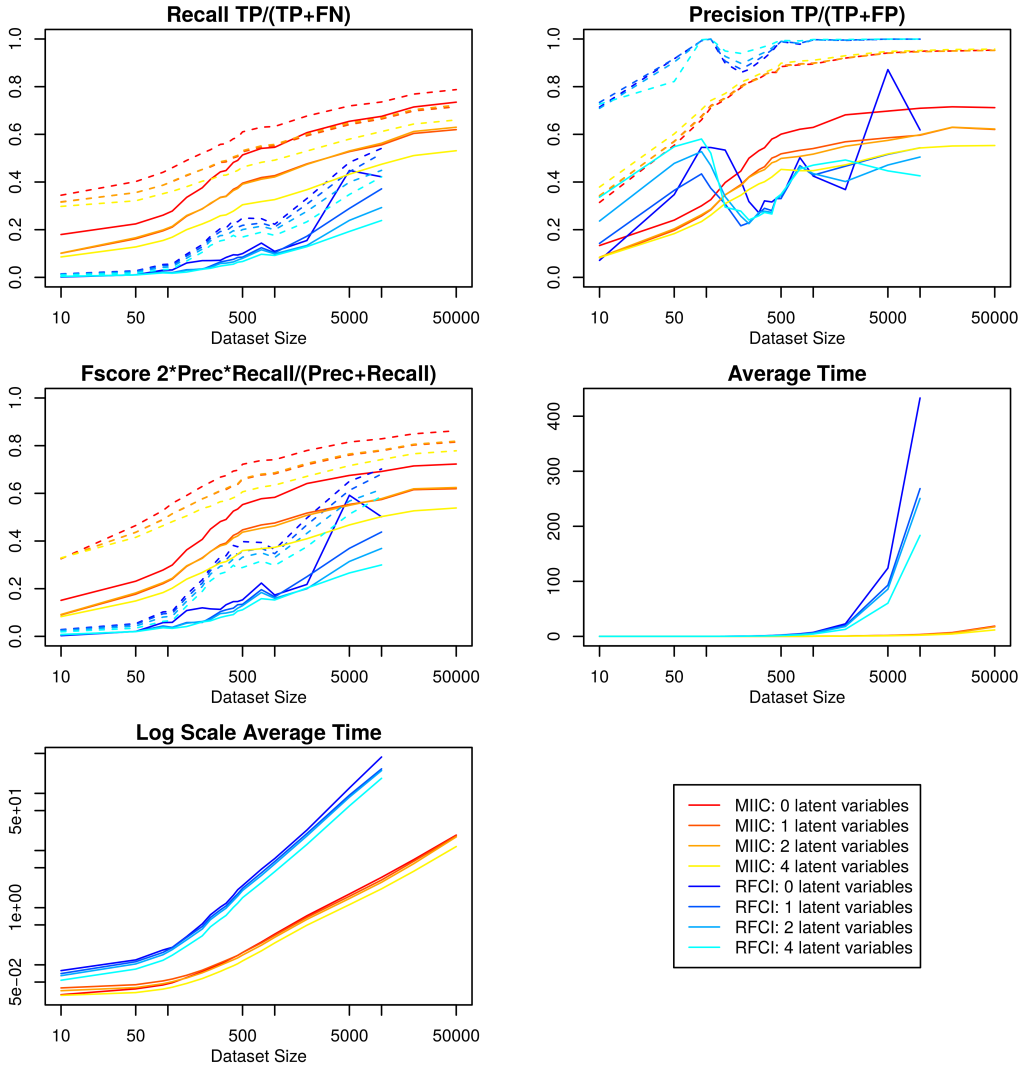



FIGURE 11.3 – **Résultats complets : réseau benchmark Insurance**

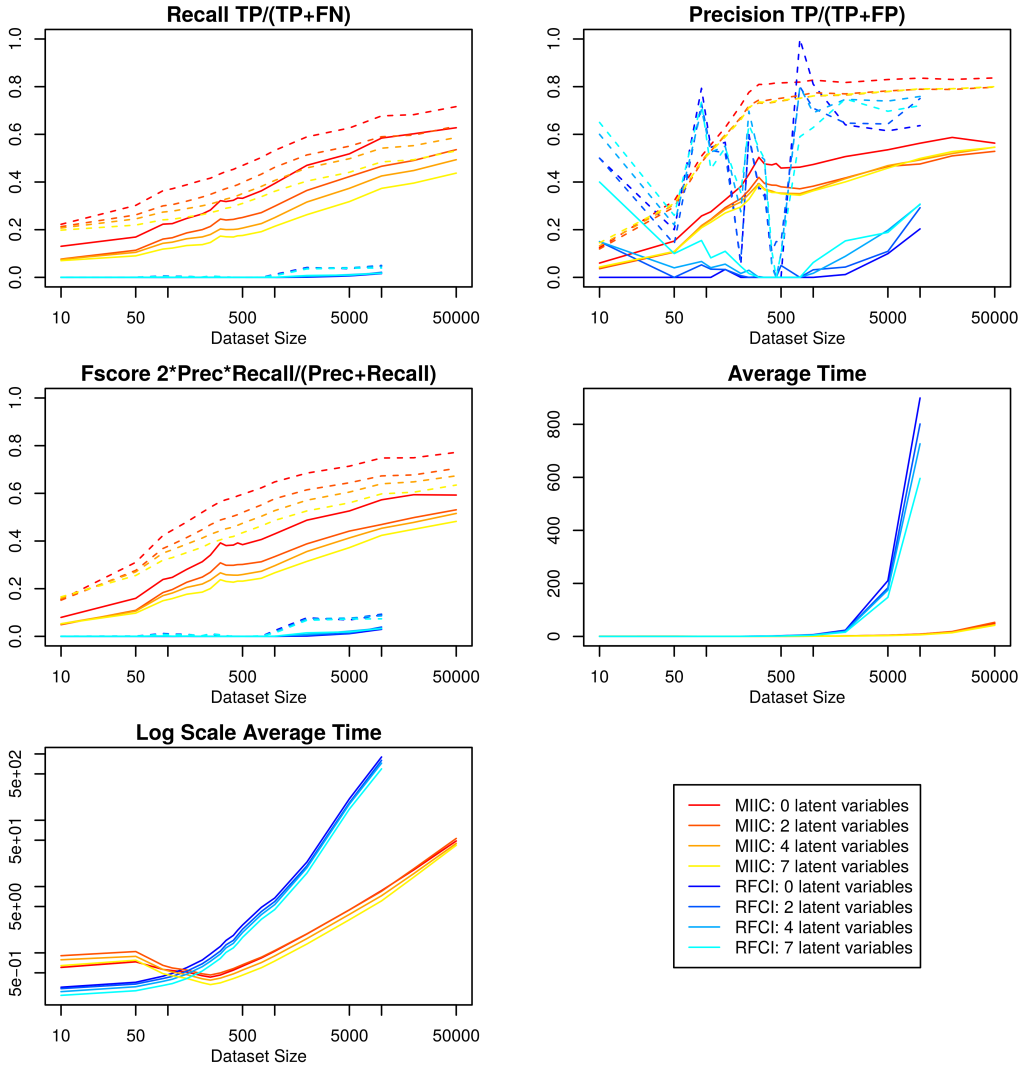



FIGURE 11.4 – **Résultats complets : réseau benchmark Barley**

# BIBLIOGRAPHIE


[1] Séverine AFFELDT. « Reconstruction de réseaux fonctionnels et analyse causale en biologie des systèmes ». 2015PA066171. Thèse de doct. 2015.

[2] Séverine AFFELDT, Louis VERNY et Hervé ISAMBERT. « 3off2 : A network reconstruction algorithm based on 2-point and 3-point information statistics ». In : *BMC Bioinformatics* **17**.*2* (2016), S12. ISSN : 1471-2105. DOI : 10.1186/s12859-015-0856-x.

[3] Hirotugu AKAIKE. « A new look at the statistical model identification ». In : *Automatic Control, IEEE Transactions on* **19**.*6* (déc. 1974), p. 716–723. ISSN : 0018-9286. DOI : 10.1109/tac.1974.1100705.

[4] Margaret H. BARON. « Concise Review : Early Embryonic Erythropoiesis : Not so Primitive After All ». In : *STEM CELLS* **31**.*5* (2013), p. 849–856. DOI : 10.1002/stem.1342.

[5] Diego COLOMBO et Marloes H. MAATHUIS. « Order-independent Constraint-based Causal Structure Learning ». In : *J. Mach. Learn. Res.* **15**.*1* (jan. 2014), p. 3741–3782. ISSN : 1532-4435.

[6] Diego COLOMBO et al. « Learning high-dimensional directed acyclic graphs with latent and selection variables ». In : *Ann. Statist.* **40**.*1* (fév. 2012), p. 294–321. DOI : 10.1214/11-AOS940.

[7] Paul C DONAGHY et Ian G MCKEITH. « The clinical characteristics of dementia with Lewy bodies and a consideration of prodromal diagnosis ». In : *Alzheimer's Research and Therapy* **6** (2014), p. 46. DOI : 10.1186/alzrt274.

[8] J. T. EPPIG et al. « The Mouse Genome Database (MGD) : comprehensive resource for genetics and genomics of the laboratory mouse ». In : *Nucleic Acids Res.* **40**.*Database issue* (jan. 2012), p. D881–886.

[9] Fazekas F et al. « MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging ». In : *American Journal of Roentgenology* **149** (1987), p. 351–356. DOI : 10.2214/ajr.149.2.351.

[10] Isabel FERREIRÓS-VIDAL et al. « Genome-wide identification of Ikaros targets elucidates its contribution to mouse B-cell lineage specification and pre-B–cell differentiation ». In : *Blood* **121**.*10* (2013), p. 1769–1782. ISSN : 0006-4971. DOI : 10.1182/blood-2012-08-450114.

[11] Cassidy M. FIFORD et al. « White matter hyperintensities are associated with disproportionate progressive hippocampal atrophy ». In : *Hippocampus* **27**.*3* (2017), p. 249–262. ISSN : 1098-1063. DOI : 10.1002/hipo.22690.

[12] N. C. FOX et al. « Presymptomatic hippocampal atrophy in Alzheimer's diseaseA longitudinal MRI study ». In : *Brain* **119**.*6* (1996), p. 2001–2007. DOI : 10.1093/brain/119.6.2001.

[13] C. GLYMOUR, P. SPIRTES et R. SCHEINES. « Causal Inference ». In : *Erkenntnis* **35**.*1-3* (1991), p. 151–189.

[14] A-S. HAMY et al. « Stromal lymphocyte infiltration after neoadjuvant chemotherapy is associated with aggressive residual disease and lower disease-free survival in HER2-positive breast cancer ». In : *Ann Oncol* (2017). DOI : 10.1093/annonc/mdx309.

[15] A.-S. HAMY-PETIT et al. « Pathological complete response and prognosis after neoadjuvant chemotherapy for HER2-positive breast cancers before and after trastuzumab era : results from a real-life cohort ». In : *British Journal of Cancer* (2016). DOI : 10.1038/bjc.2015.426.

[16] J. A. HARTIGAN et M. A. WONG. « A k-means clustering algorithm ». In : *JSTOR : Applied Statistics* **28**.*1* (1979), p. 100–108.

[17] David HECKERMAN, Dan GEIGER et David M. CHICKERING. « Learning Bayesian Networks : The Combination of Knowledge and Statistical Data ». In : *Machine Learning* **20**.*3* (sept. 1995), p. 197–243. ISSN : 1573-0565. DOI : 10.1023/A:1022623210503.

[18] Michael I. JORDAN, éd. *Learning in Graphical Models*. Cambridge, MA, USA : MIT Press, 1999. ISBN : 0-262-60032-3.

[19] A. KAKIZUKA et al. « Chromosomal translocation t(15 ;17) in human acute promyelocytic leukemia fuses RAR alpha with a novel putative transcription factor, PML ». In : *Cell* **66**.*4* (1991), p. 663–674. ISSN : 0092-8674. DOI : http://dx.doi.org/10.1016/0092-8674(91)90112-C.

[20] Tonáš KOČKA et Robert CASTELO. « Improved Learning of Bayesian Networks ». In : *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. UAI'01. Seattle, Washington : Morgan Kaufmann Publishers Inc., 2001, p. 269–276. ISBN : 1-55860-800-1.

[21] Christophe LANCRIN et al. « The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage ». In : *Nature* **457**.*4* (2009), p. 892–895. ISSN : 0028-0836. DOI : http://dx.doi.org/10.1038/nature07679.

[22] Carlos O. LIZAMA et al. « Repression of arterial genes in hemogenic endothelium is sufficient for haematopoietic fate acquisition ». In : *Nature Communications* **6** (2015), p. 7739. DOI : http://dx.doi.org/10.1038/ncomms8739.

[23] LOOK et A. THOMAS. « Oncogenic Transcription Factors in the Human Acute Leukemias ». In : *Science* **278**.*5340* (1997), p. 1059–1064. DOI : 10.1126/science.278.5340.1059.

[24] KA MATTHEWS et al. « Body mass index in mid-life women : relative influence of menopause, hormone use, and ethnicity. » In : *Int J Obes Relat Metab Disord* (2001). DOI : 10.1038/sj.ijo.0801618.

[25] Victoria MOIGNARD et al. « Decoding the regulatory network of early blood development from single-cell gene expression measurements ». In : *Nat Biotechnol* **33**.*3* (2015), p. 269–276. DOI : 10.1038/nbt.3154.

[26] Victoria MOIGNARD et al. « Transcriptional hierarchies regulating early blood cell development ». In : *Blood Cells, Molecules, and Diseases* **51**.*4* (2013). Embryonic and Fetal Hematopoiesis, p. 239–247. ISSN : 1079-9796. DOI : http://doi.org/10.1016/j.bcmd.2013.07.007.

[27] James PALIS et al. « Primitive erythropoiesis in the mammalian embryo ». In : *Int. J. Dev. Biol* **54** (2010), p. 1011–1018.

[28] Nikolaos D. PAPATHANASIOU et al. « Diagnostic accuracy of 123I-FP-CIT (DaTSCAN) in dementia with Lewy bodies : A meta-analysis of published studies ». In : *Parkinsonism and Related Disorders* **18**.*3* (2012), p. 225–229. ISSN : 1353-8020. DOI : https://doi.org/10.1016/j.parkreldis.2011.09.015.

[29] Judea PEARL. *Causality : Models, Reasoning, and Inference*. New York, NY, USA : Cambridge University Press, 2000. ISBN : 0-521-77362-8.

[30] Judea PEARL. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1988. ISBN : 0-934613-73-7.

[31] John E. PIMANDA et al. « Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development ». In : *Proceedings of the National Academy of Sciences* **104**.*45* (2007), p. 17692–17697. DOI : 10.1073/pnas.0707045104.

[32] Niels D PRINS et Philip SCHELTENS. « White matter hyperintensities, cognitive impairment and dementia : an update ». In : *Nature Reviews Neurology* **11**.*5* (2015), p. 157–165. DOI : 10.1038/nrneurol.2015.10.

[33] Joseph RAMSEY, Jiji ZHANG et Peter SPIRTES. « Adjacency-Faithfulness and Conservative Causal Inference ». In : *UAI '06, Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, July 13-16, 2006*. 2006.

[34] Thomas RICHARDSON et Peter SPIRTES. « Ancestral Graph Markov Models ». In : *Annals of Statistics* **30** (2000), p. 2002.

[35] Baulies S. et al. « Time-varying effect. and long-term survival analysis in breast cancer patients treated with neoadjuvant chemotherapy ». In : *Br J Cancer* **113** (2015). ISSN : 0007-0920. DOI : 10.1038/bjc.2015.174.

[36] P SCHELTENS et al. « Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing : diagnostic value and neuropsychological correlates. » In : *Journal of Neurology, Neurosurgery & Psychiatry* **55**.*10* (1992), p. 967–972. ISSN : 0022-3050. DOI : 10.1136/jnnp.55.10.967. eprint : http://jnnp.bmj.com/content/55/10/967.full.pdf.

[37] Gideon SCHWARZ. « Estimating the Dimension of a Model ». In : *Ann. Statist.* **6**.*2* (mar. 1978), p. 461–464. DOI : 10.1214/aos/1176344136.

[38] J M SENARD et al. « Prevalence of orthostatic hypotension in Parkinson's disease ». In : *Journal of Neurology, Neurosurgery & Psychiatry* **63**.*5* (1997), p. 584–589. ISSN : 0022-3050. DOI : 10.1136/jnnp.63.5.584. eprint : http://jnnp.bmj.com/content/63/5/584.full.pdf.

[39] Y I SHELINE et al. « Hippocampal atrophy in recurrent major depression ». In : *Proceedings of the National Academy of Sciences* **93**.*9* (1996), p. 3908–3913.

[40] *Brain Aging, Biological and Psychosocial Aspect*. T. 189. 7. 2005, p. 1371–1381.

[41] P. SPIRTES, C. GLYMOUR et R. SCHEINES. *Causation, Prediction, and Search*. 2nd. MIT press, 2000.

[42] Peter SPIRTES, Christopher MEEK et Thomas RICHARDSON. « Causal Inference in the Presence of Latent Variables and Selection Bias ». In : *In Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, p. 499–506.

[43] Moignard V. et al. « Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis ». In : *Nat Cell Biol* **15**.*5* (2013). DOI : 10.1038/ncb2709.

[44] Adam C. WILKINSON et Berthold GÖTTGENS. « Transcriptional Regulation of Haematopoietic Stem Cells ». In : *Transcriptional and Translational Regulation of Stem Cells*. Sous la dir. de Gary HIME et Helen ABUD. Springer Netherlands, 2013, p. 187–212. ISBN : 978-94-007-6621-1. DOI : 10.1007/978-94-007-6621-1_11.

[45] Jiji ZHANG. « Causal Reasoning with Ancestral Graphs ». In : *Journal of Machine Learning Research* **9** (2008), p. 1437–1474. DOI : 10.1145/1390681.1442780.

[46] Jiji ZHANG. « On the Completeness of Orientation Rules for Causal Discovery in the Presence of Latent Confounders and Selection Bias ». In : *Artif. Intell.* **172**.*16-17* (nov. 2008), p. 1873–1896. ISSN : 0004-3702. DOI : 10.1016/j.artint.2008.08.001.