The Bayesys data and Bayesian network repository

Anthony C. Constantinou[1,2], Yang Liu[1], Kiattikun Chobtham[1],
Zhigao Guo[1], and Neville K. Kitson[1]

Version 1.8[1]
(last revision: April 2025)

a) Bayesian AI research lab,
Machine Intelligence and Decision Systems (MInDS) research group,
School of Electronic Engineering and Computer Science,
Queen Mary University of London,
London, UK, E1 4FZ.

E-mail: a.constantinou@qmul.ac.uk

b) The Alan Turing Institute, UK, British Library, London, UK, NW1 2DB.

MInDS
Machine Intelligence and Decision Systems
Research Group

Bayesian Artificial Intelligence
Research Lab

Queen Mary
University of London

The
Alan Turing
Institute

Engineering and Physical Sciences
Research Council

BAYESFUSION,LLC
Data Analytics, Modeling, Decision Support

www.bayesfusion.com

agena.ai

www.agena.ai

[1] **Citation:** Constantinou, A. C., Liu, Y., Chobtham, K., Guo, Z., and Kitson, N. K. (2020). The Bayesys data and Bayesian network repository. Bayesian AI research lab, MInDS research group, Queen Mary University of London, London, UK. [Online]. Available: http://bayesian-ai.eecs.qmul.ac.uk/bayesys/

Table of Contents

Copyright notice

Acknowledgements

This work was partially supported by the ERSRC Fellowship project EP/S001646/1 "*Bayesian Artificial Intelligence for Decision Making under Uncertainty*" [1], by the project partner Agena Ltd, and by The Alan Turing Institute in the UK under the EPSRC grant EP/N510129/1.

1. Algorithm implementations

The section lists the algorithms we have implemented listed earliest implementation. Each algorithm is available either as a standalone package or in Bayesys:

1) **SaiyanH**: a hybrid structure learning algorithm described in [2], which is based on an earlier version of this algorithm described in [3]. It starts with a dependency function that produces an undirected graph that can be viewed as an extended version of the maximum spanning graph, ensuring that each variable associates with at least one edge. It then obtains a DAG using a combination of constraint-based, score-based and interventional impact rules. The DAG is then provided as input to Tabu search, with the restriction not to delete edges that lead to disjoint subgraphs or disjoint nodes. The algorithm is available in Bayesys [4].

2) **CCHM:** A hybrid structure learning algorithm that combines the constraint-based part of cFCI with hill-climbing score-based learning and Pearl's do-operator to measure causal effects. CCHM assumes the presence of latent variables and that the BN is a linear structure equation model where data follow a multivariate Gaussian distribution. The algorithm is described in [5] and is available as a standalone package here: https://github.com/kiattikunc/CCHM .

3) **SED:** A score-based algorithm that can be added as an additional learning phase at the end of any structure learning algorithm, and serves as a correction learning phase that removes potential false positive edges produced by other algorithms, often due to measurement error. The algorithm is described in [6] and is available as a standalone package here: https://github.com/Enderlogic/Spurious-Edge-Detection

4) **HC:** Represents the classic score-based hill-climbing search. It starts from an empty graph and performs hill-climbing search by exploring arc additions, reversals and deletions, moving to the DAG that maximises the BIC score. The algorithm is available in Bayesys [4].

5) **TABU:** The classic tabu search structure learning algorithm. It assumes discrete variables as input. It starts from an empty graph and moves to the neighbouring DAG that maximises the BIC score. When it reaches a local maximum, it performs $|V|$ escapes (where V is the set of variables in the data) that minimally decrease the BIC score, and repeats hill-climbing on each iteration of $|V|$ in an attempt to escape a local maximum. The algorithm is available in Bayesys [4].

6) **HC-aIPW:** A greedy search structure learning algorithm suitable for discrete datasets that contain missing values that are missing at random and not at random. The algorithm utilises pairwise deletion and inverse probability weighting to maximally leverage the observed data and to limit potential bias caused by missing values. The algorithm is described in [7] and is available as a standalone package here: https://github.com/Enderlogic/HC-missing-data

7) **mFGS-BS:** A hybrid algorithm for structure learning from discrete observational and interventional data. The algorithm assumes causal insufficiency in the presence of latent variables and produces a Partial Ancestral Graph (PAG) output. It relies on a hybrid learning approach and a novel Bayesian scoring paradigm that calculates the posterior probability of each directed edge being added to the learnt graph. The algorithm is described in [8] and is available as a standalone package here: https://github.com/kiattikunc/mFGES-BS

8) **MAHC:** The Model-Averaging Hill-Climbing is a score-based algorithm described in [9]. It combines two novel strategies with hill-climbing search. The algorithm starts by pruning the search space of graphs, where the pruning strategy can be viewed as an aggressive version of the pruning strategies that are typically applied to combinatorial optimisation structure learning problems. It then performs model averaging in the hill-climbing search process and moves to the neighbouring graph that maximises the objective function, on average, for that neighbour and over all its valid neighbours. The MAHC algorithm is available in Bayesys [4].

9) **PS-MINOBS:** A score-based algorithm suitable for high dimensional data that can be viewed as an extension of MINOBS. The algorithm constructs the graph by sampling Candidate Parent Sets (CPSs) for each variable. Sampling is performed in parallel under the assumption the distribution of CPSs is half-normal and ordered by Bayesian score for each variable. Sampling from a half-normal distribution ensures that the CPSs sampled are likely to be those which produce the higher scores. The algorithm is described in [10] and is available as a standalone package here: https://github.com/ZHIGAO-GUO/Parallel-MINOBS

10) **ILC-V** and **HCLC-V:** Two hybrid learning algorithms for discovery and density estimation of latent confounders in causal structural model learning. They combine elements of variational Bayesian methods, expectation-maximisation, hill-climbing search, and structure learning under the assumption of causal insufficiency. ILC-V maximises model selection accuracy, whereas HCLC-V improves computational efficiency in exchange for minor reductions in accuracy. The former algorithm is suitable for small networks and the latter for moderate size networks. The algorithms are described in [11] and are available as a standalone package here: https://github.com/kiattikunc/ILC_and_HCLC_with_VBEM

11) **MBMF:** The Markov-Blanket MissForest recovers the Markov blanket of partially observed variables based on the graphical expression of missingness known as the m-graph, which captures observed variables together with missingness indicators and the possible causal links between them. It combined the Markov blanket approach with MissForest, to formulate a new algorithm that improves imputation accuracy under both random and systematic missingness. This algorithm is described in [24], and is available as a standalone package here: https://github.com/Enderlogic/Markov-Blanket-based-Feature-Selection

2. Case study networks

2.01. ASIA network

A BN for diagnosing patients at a clinic [13].

- Download the ground truth DAG in CSV format.
- Download synthetic data generated from the ground truth DAG in CSV format (100,000 samples).

Table 2.01. The properties of the ground truth ASIA DAG.

Number of variables:	8
Number of edges:	8
Number of free parameters:	18
Maximum in-degree:	2
Maximum out-degree:	2
Maximum degree:	4

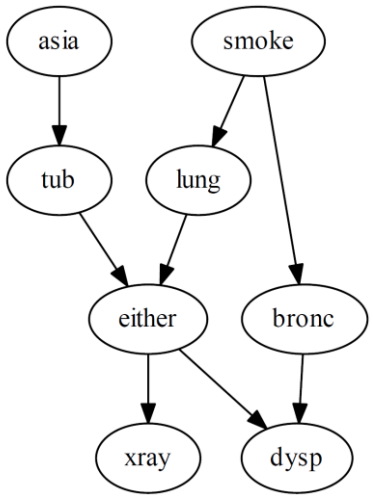

Figure 2.01. The ground truth DAG of the ASIA network with 8 nodes and 8 directed edges.

2.02. ALARM network

A BN based on an alarm message system for patient monitoring [12]:

- Download the ground truth DAG in CSV format.
- Download synthetic data generated from the ground truth DAG in CSV format (100,000 samples).

Table 2.02. The properties of the ground truth ALARM DAG.

Number of variables:	37
Number of edges:	46
Number of free parameters:	509
Maximum in-degree:	4
Maximum out-degree:	5
Maximum degree:	6

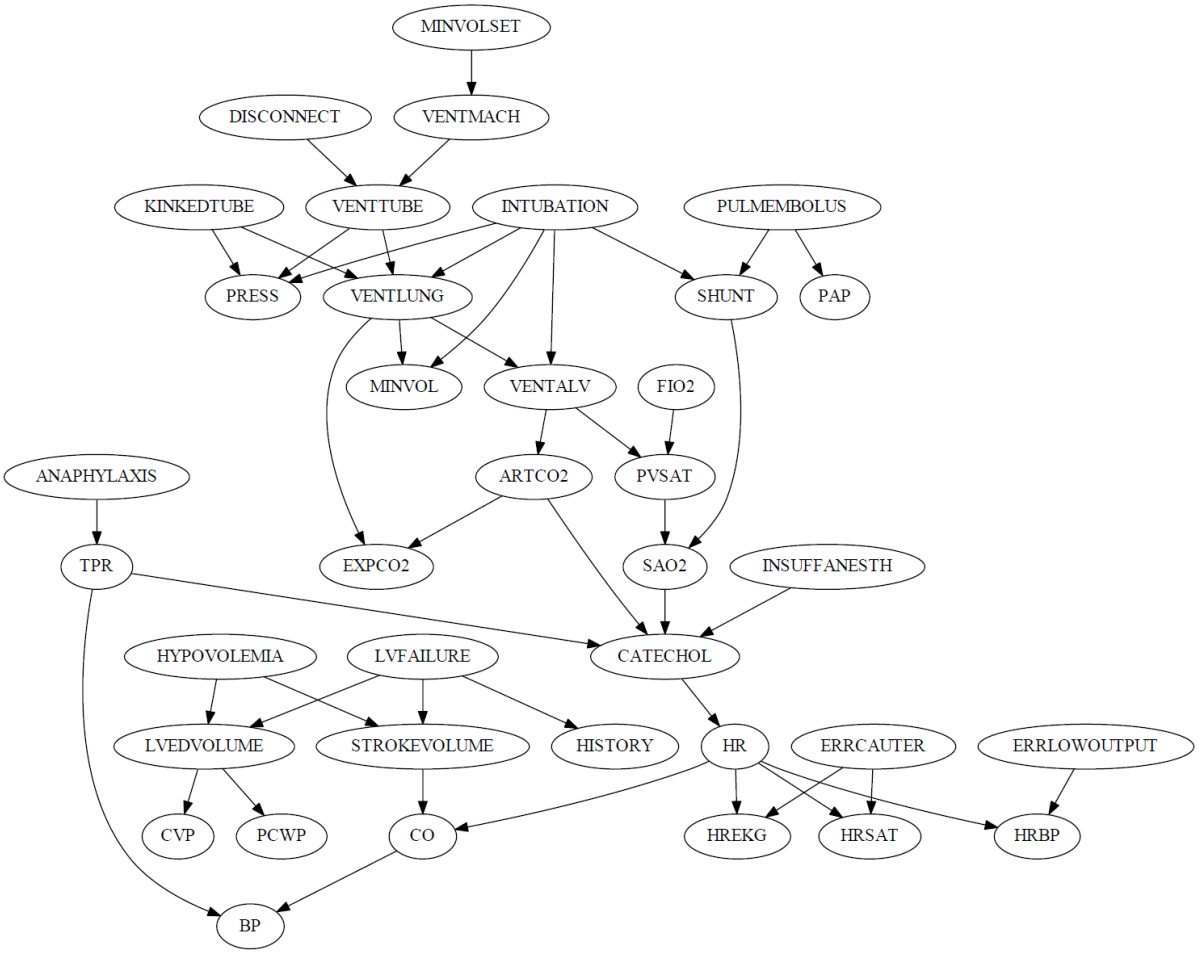

Figure 2.02. The ground truth DAG of the ALARM network with 37 nodes and 46 directed edges.

2.03. DIARRHOEA network (with access to real data)

A BN for modelling the risk of childhood diarrhoea in India [14].

- Download the real dataset (259,627 samples). Missing values were imputed using the standard MissForest imputation algorithm.
- Download the ground truth DAG in CSV format.
- Download synthetic data generated from the ground truth DAG in CSV format (100,000 samples).

Table 2.03. The properties of the ground truth DIARRHOEA DAG.

Number of variables:	28
Number of edges:	68
Number of free parameters:	1,716
Maximum in-degree:	8
Maximum out-degree:	15
Maximum degree:	17

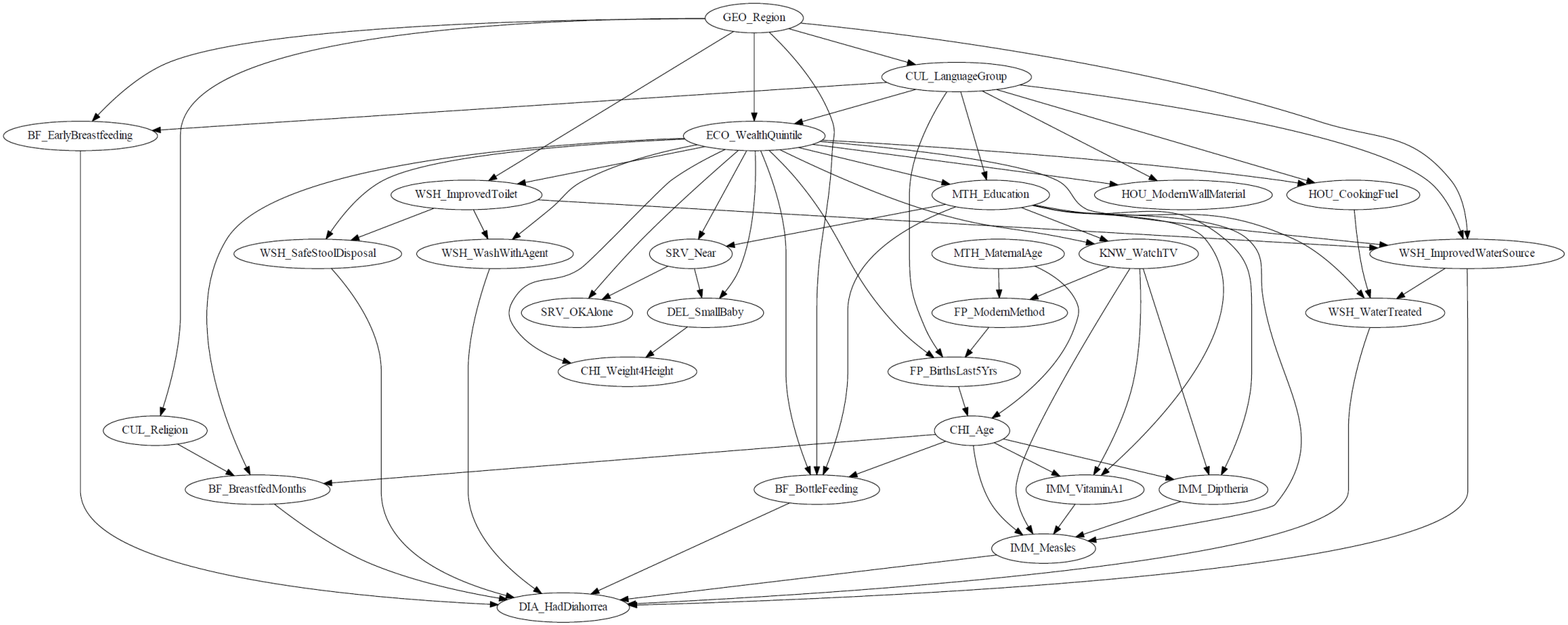

Figure 2.03. The ground truth DAG of the DIARRHOEA network with 28 nodes and 68 edges.

2.04. FORMED network (with access to real data)

A BN that captures the risk of violent reoffending of mentally ill prisoners, along with multiple interventions for managing this risk [15]:

- Download the real dataset (953 samples). The data value 'unknown' indicates missing value.
- Download the ground truth DAG_1 in CSV format, based on the variables present in the real dataset above. Note this graph excludes the synthetic variables described in [15], and this makes DAG_1 less realistic than the DAG_2 below.
- Download the ground truth DAG_2 in CSV format, based on the variables present in the real dataset above, plus the synthetic variables described in [15].
- Download synthetic data generated from the ground truth DAG_2 in CSV format (100,000 samples).

Table 2.04. The properties of the two ground truth FORMED DAGs.

	DAG 1	DAG 2
Number of variables:	56	88
Number of edges:	95	138
Number of free parameters:	39,196,846	912
Maximum in-degree:	17	6
Maximum out-degree:	11	10
Maximum degree:	22	11

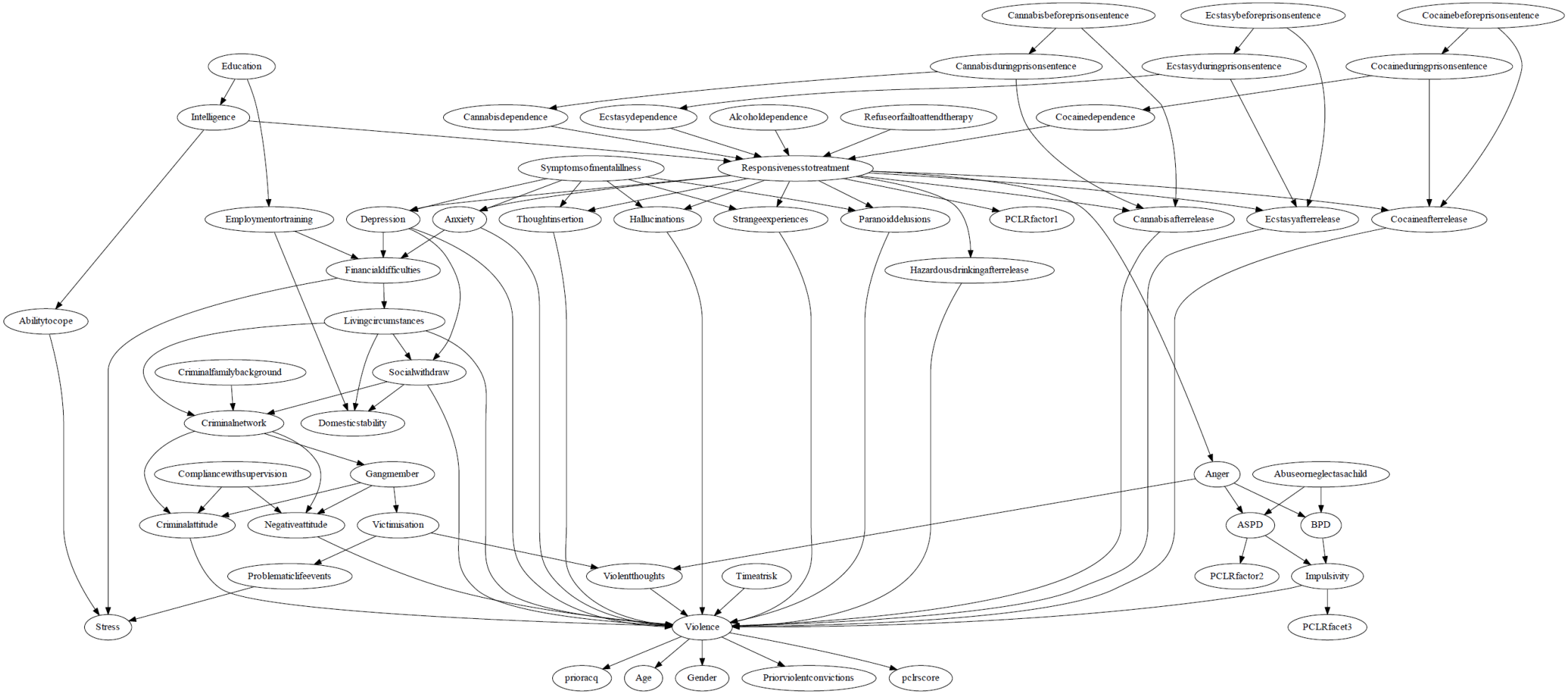

Figure 2.041. The ground truth DAG_1 of the FORMED network with 56 nodes and 95 edges.

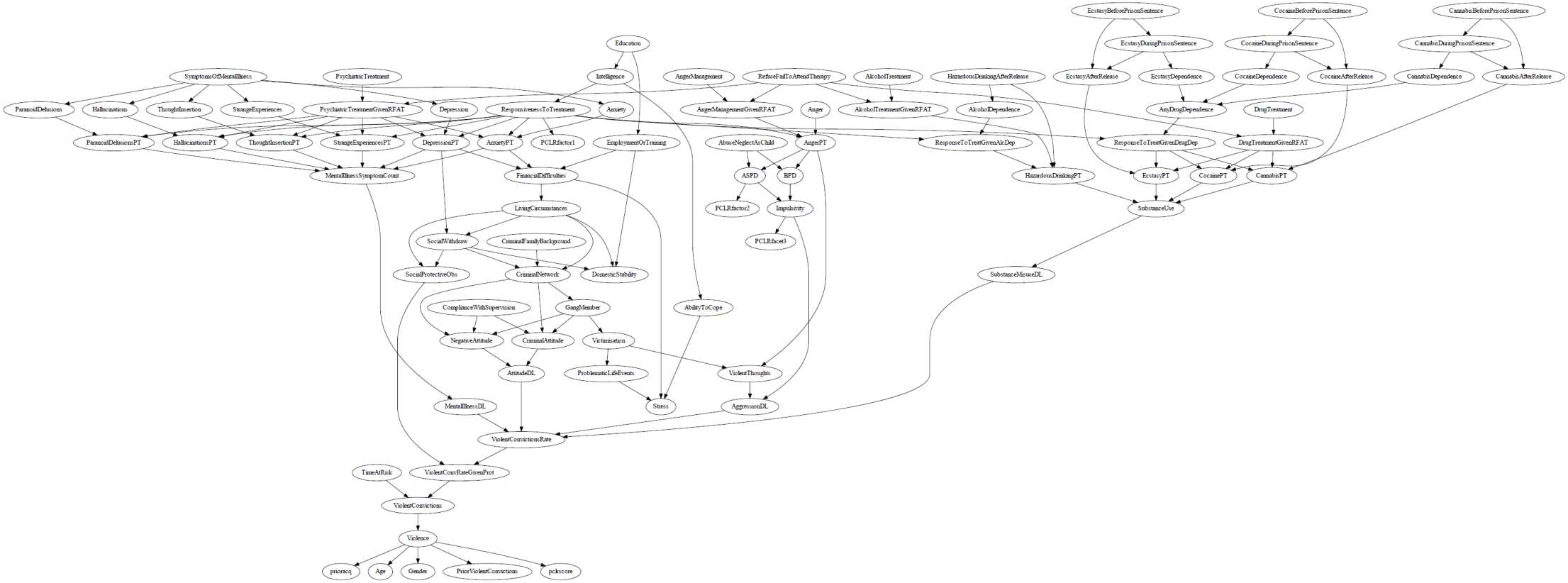

Figure 2.042. The ground truth DAG_2 of the FORMED network with 88 nodes and 138 edges.

2.05. PATHFINDER network

A BN that was designed to assist surgical pathologists with the diagnosis of lymph-node diseases [16]:

- Download the ground truth DAG in CSV format.
- Download synthetic data generated from the ground truth DAG in CSV format (100,000 samples).

Table 2.05. The properties of the ground truth PATHFINDER DAG.

Number of variables:	109
Number of edges:	195
Number of free parameters:	71,890
Maximum in-degree:	5
Maximum out-degree:	106
Maximum degree:	106

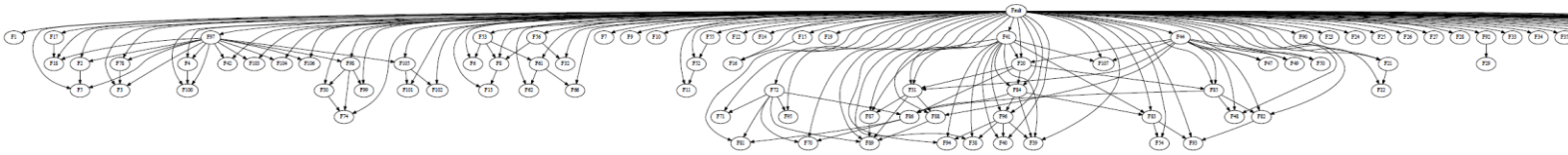

Figure 2.05. The ground truth DAG of the PATHFINDER network with 109 nodes and 195 edges.

2.6. PROPERTY network

A BN that assesses investment decisions in the UK property market [17]:

- Download the ground truth DAG in CSV format.
- Download synthetic data generated from the ground truth DAG in CSV format (100,000 samples).

Table 2.06. The properties of the ground truth PROPERTY DAG.

Number of variables:	27
Number of edges:	31
Number of free parameters:	3,056
Maximum in-degree:	3
Maximum out-degree:	6
Maximum degree:	6

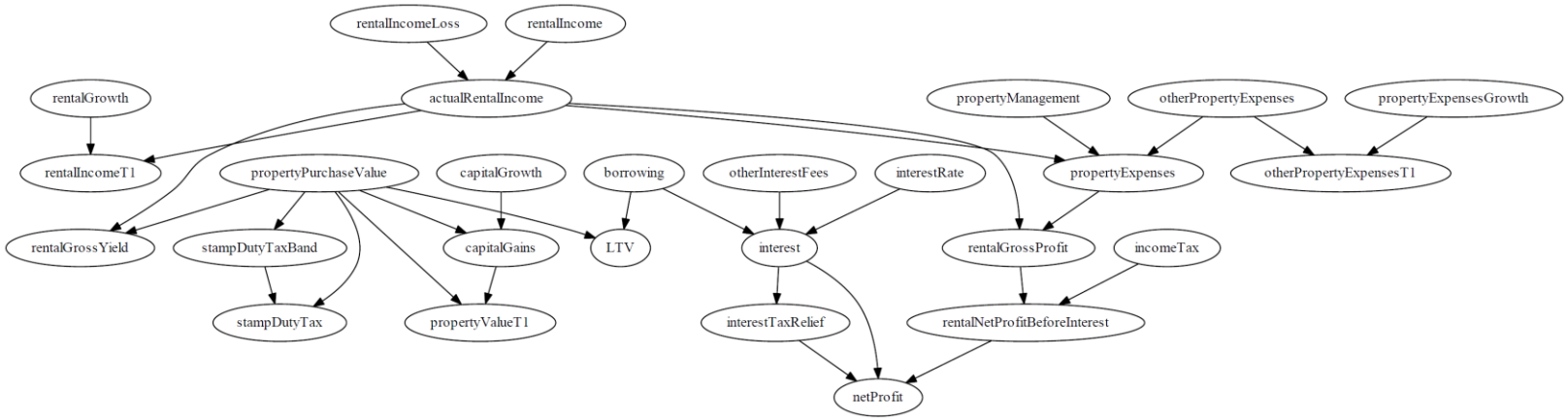

Figure 2.06. The ground truth DAG of the PROPERTY network with 27 nodes and 31 edges.

2.07. SPORTS network (with access to real data)

A BN that combines football team ratings with various team performance statistics to predict various match score outcomes [18]:

- Download the real dataset (3536 samples)
- Download the ground truth DAG in CSV format.
- Download synthetic data generated from the ground truth DAG in CSV format (100,000 samples).

Table 2.07. The properties of the ground truth SPORTS DAG.

Number of variables:	9
Number of edges:	15
Number of free parameters:	1,049
Maximum in-degree:	2
Maximum out-degree:	7
Maximum degree:	7

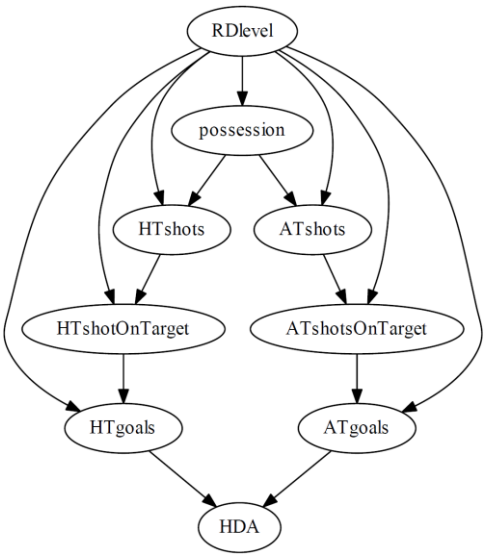

Figure 2.07. The ground truth DAG of the SPORTS network with 9 nodes and 15 directed edges.

2.08. COVID-19 network (with access to real data)

A BN that captures the causal structure of the COVID-19 pandemic in the UK [19]:

- Download the real dataset. There are four datasets, as described in [19]: a) the raw mixed dataset, b) the corresponding continuous dataset, c) the dataset discretised with k-means clustering, d) the dataset discretised with quartiles (866 samples). Missing values are imputed using the Markov-Blanket MissForest (MBMF) imputation algorithm [24].
- Download the ground truth DAG in CSV format.
- Download synthetic data generated from the ground truth DAG in CSV format (100,000 samples).

Table 2.08. The properties of the ground truth COVID-19 DAG.

Number of variables:	17
Number of edges:	37
Number of free parameters:	7,834
Maximum in-degree:	5
Maximum out-degree:	7
Maximum degree:	10

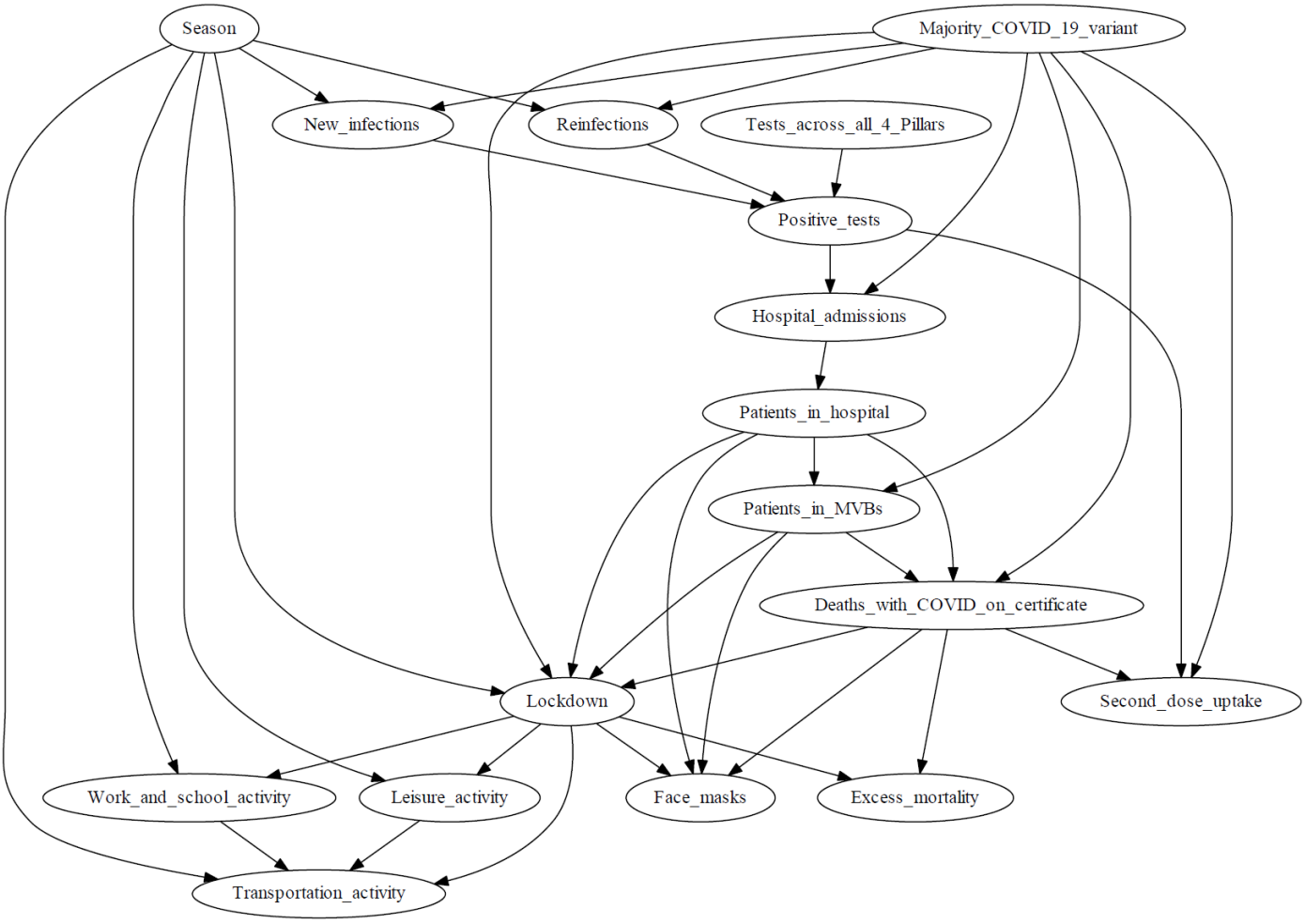

Figure 2.08. The ground truth DAG of the COVID-19 network, with 17 nodes and 37 directed edges.

2.09. DIABETES network (with access to real data)

A BN that captures causal pathways amongst potential risk factors influencing diabetes progression [20]:

- Download the real dataset, containing 22 variables and 253,680 samples.
- Download the ground truth DAGs in CSV format. This ZIP file includes three structures elicited by a domain expert who categorised graph edges into three levels of confidence (low, moderate, and high), leading into three individual graphs corresponding to each of those levels of confidence.
- Download the Bayesian network of the high-confidence ground truth DAG in GeNIe format.
- Download the Bayesian network of the model-averaging learnt DAG, obtained over multiple structure learning algorithms as specified in [20], in GeNIe format.
- Download synthetic data generated from the high-confidence ground truth DAG in CSV format (100,000 samples).

Table 2.09. The properties of the high-confidence ground truth DIABETES DAG.

Number of variables:	22
Number of edges:	37
Number of free parameters:	1,348
Maximum in-degree:	7
Maximum out-degree:	7
Maximum degree:	8

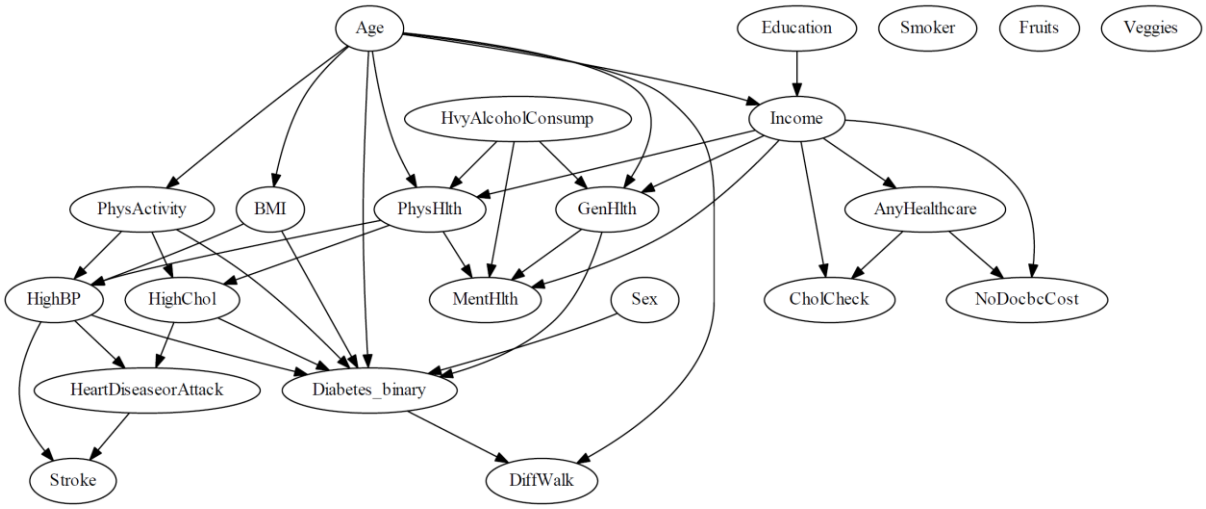

Figure 2.59. The ground truth DAG of the DIABETES high-confidence network with 22 nodes and 37 directed edges.

2.10 Gestational Diabetes Mellitus (GDM) network.

A BN based on a novel clinical dataset from Cork University Maternity Hospital (CUMH), Ireland, covering the years 2014 to 2016 and 2020, used to explore key relationships within GDM-related data for the effective management of GDM in reducing associated risks and improving outcomes for both mother and child [21]:

- Download the model-averaging DAG in CSV format, obtained over multiple structure learning algorithms as specified in [21].
- Download the BN model of the model-averaging DAG in GeNIe format.
- Download synthetic data generated from the model-averaging DAG in CSV format (100,000 samples).

Table 2.10. The properties of the model-averaging GDM DAG.

Number of variables:	61
Number of edges:	75
Number of free parameters:	1,315
Maximum in-degree:	3
Maximum out-degree:	6
Maximum degree:	8

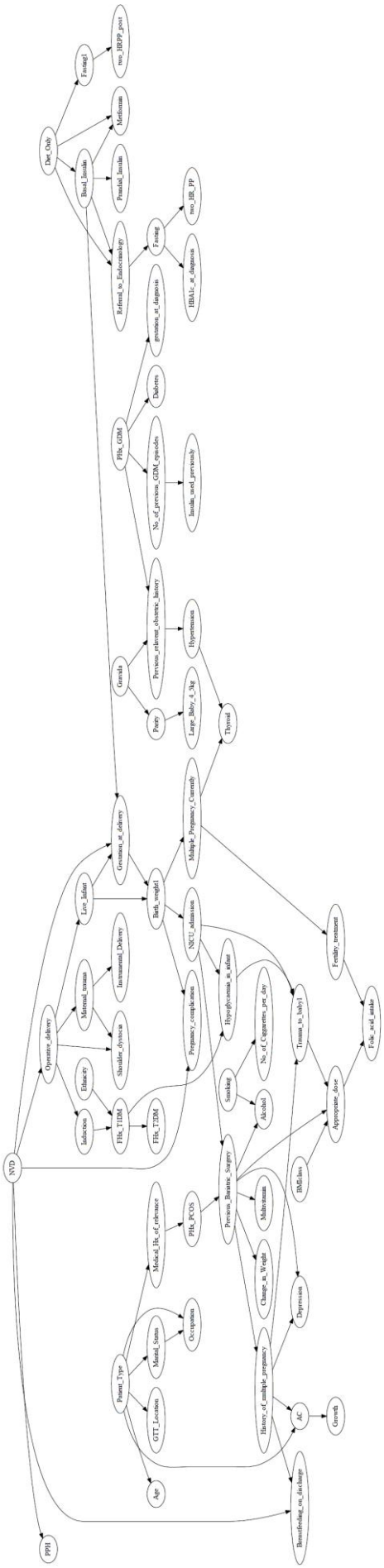

Figure 2.6. The model-averaging DAG of the GDM network, with 61 nodes and 75 directed edges.

2.11 Hattrick football manager (Hattrick) network (with access to real data)

A BN based on a novel clinical dataset from an Irish maternity hospital, covering the years 2014 to 2016 and 2020, used to explore key relationships within GDM-related data for the effective management of GDM in reducing associated risks and improving outcomes for both mother and child [22]:

- Download the discrete real dataset in CSV format, containing 250 variables and 1,000,000 samples.
- Download the continuous (includes integer variables) real dataset in CSV format, containing 250 variables and 1,000,000 samples.
- Download the knowledge-based DAG in CSV format.
- Download the discrete BN model of the knowledge-based DAG in GeNIe format.
- Download the KB_probabilistic hybrid BN model in GeNIe format, that is based on the knowledge-based DAG, but includes additional variables (363 in total) as described in Section 4 in [22]

Table 2.11. The properties of the knowledge-based DAG.

Number of variables:	250
Number of edges:	386
Number of free parameters:	483,349
Maximum in-degree:	5
Maximum out-degree:	10
Maximum degree:	13

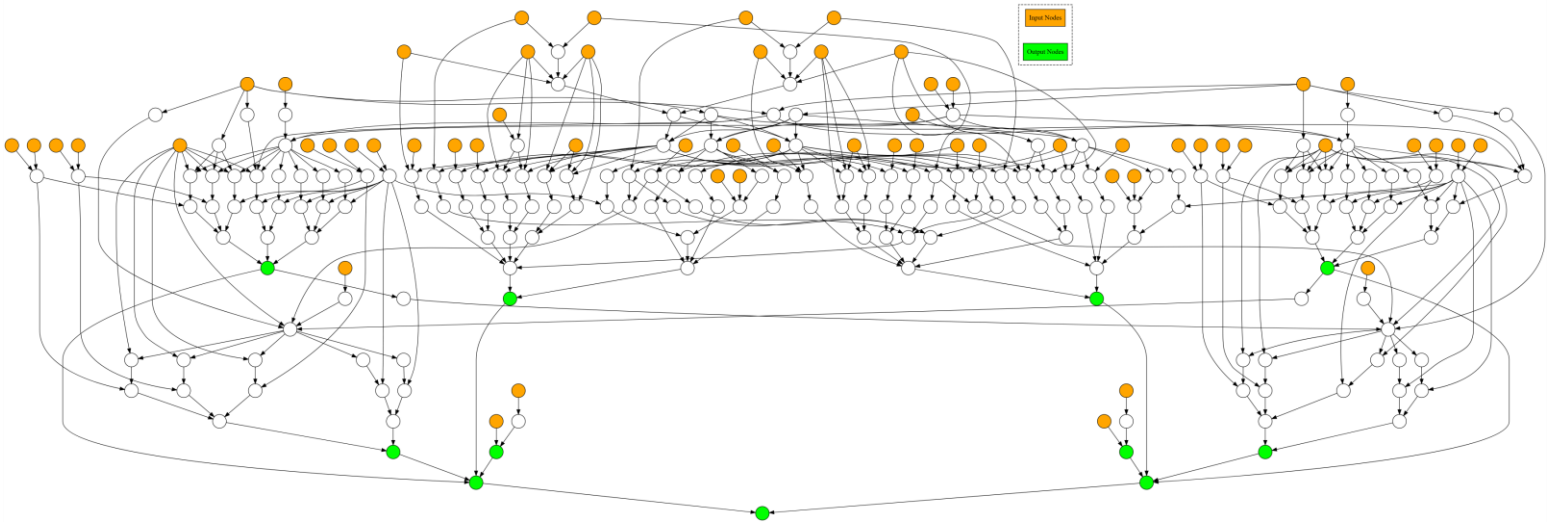

Figure 2.110. The knowledge-based DAG with input (observed) nodes in orange and output (inferred) nodes of interest in green, taken from [22].

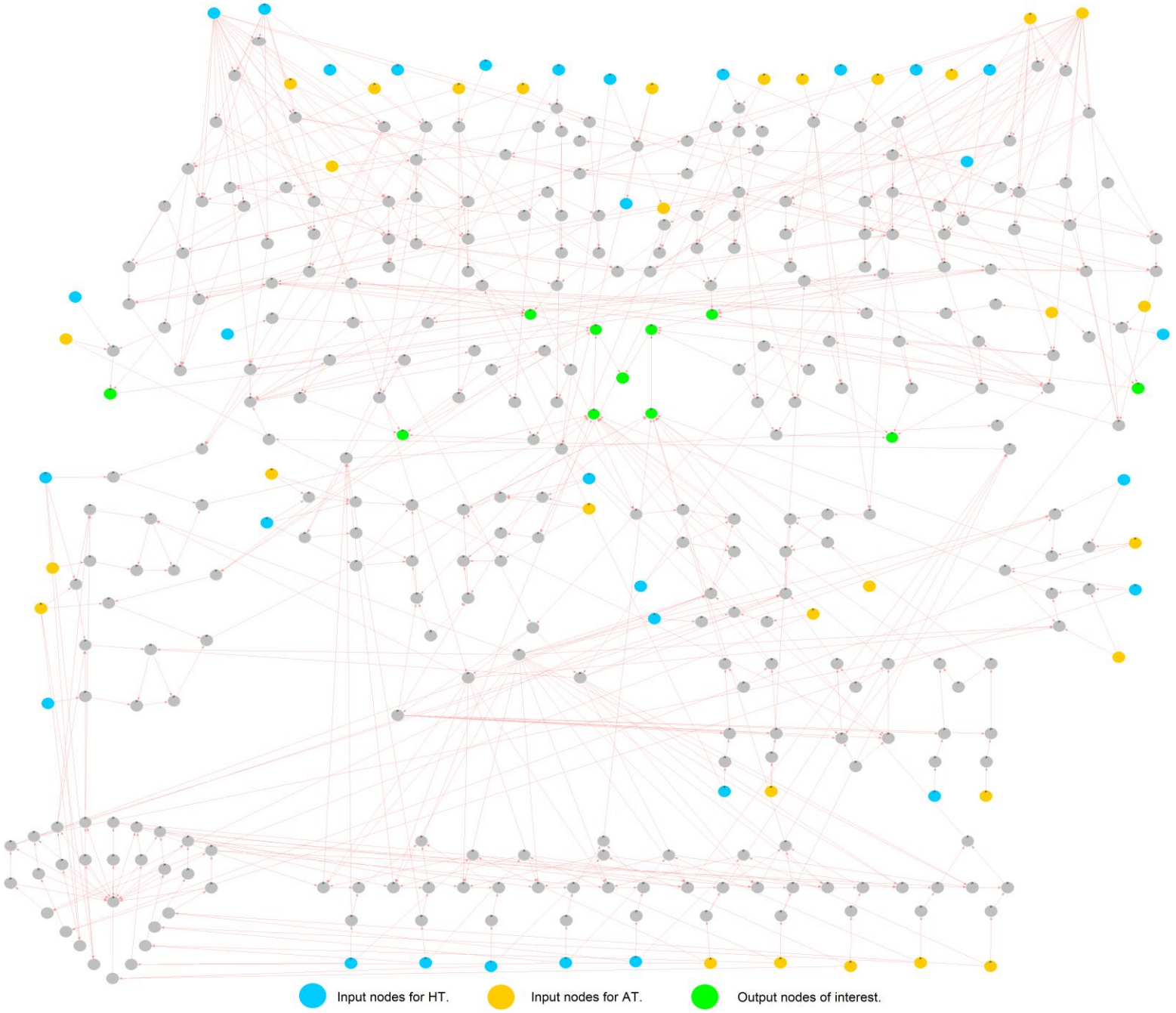

Fig. 2.111. A visualisation of the KB-probabilistic BN model, taken from [22].

Access to additional data, results and models by publication

3.1. Large-scale empirical validation of Bayesian Network structure learning algorithms with noisy data (2021)

The files available to download in this subsection come from the work in [23].

- Download all raw structure learning accuracy scores.
- Links to the synthetic datasets and the true graphs used in this study can be found in tables **Table 0.1** to **Table 0.6** in subsections 3.1.1 to 3.1.6 respectively.

3.1.1. ALARM network (clean and noisy data)

Table 0.1. The clean and noisy synthetic datasets used for the ALRAM network. Each dataset consists of five different sample sizes (from 10^2 to 10^6 samples). For details refer to [23].

Experiment no.	Experiment	Dataset (CSV)	True graph (CSV)	Notes
1	N	Link	DAG (Fig 1)	No noise
2	M5	Link	DAG (Fig 1)	Missing data (5%)
3	M10	Link	DAG (Fig 1)	Missing data (10%)
4	I5	Link	DAG (Fig 1)	Incorrect data (5%)
5	I10	Link	DAG (Fig 1)	Incorrect data (10%)
6	S5	Link	DAG (Fig 1)	Merged states data (5%)
7	S10	Link	DAG (Fig 1)	Merged states data (10%)
8	L5	Link	MAG-5 (Fig 15)	Latent confounders (5%)
9	L10	Link	MAG-10 (Fig 16)	Latent confounders (10%)
10	cMI	Link	DAG (Fig 1)	M5 and I5
11	cMS	Link	DAG (Fig 1)	M5 and S5
12	cML	Link	MAG-5 (Fig 15)	M5 and L5
13	cIS	Link	DAG (Fig 1)	I5 and S5
14	cIL	Link	MAG-5 (Fig 15)	I5 and L5
15	cSL	Link	MAG-5 (Fig 15)	S5 and L5
16	cMISL	Link	MAG-5 (Fig 15)	M5, I5, S5, and L5

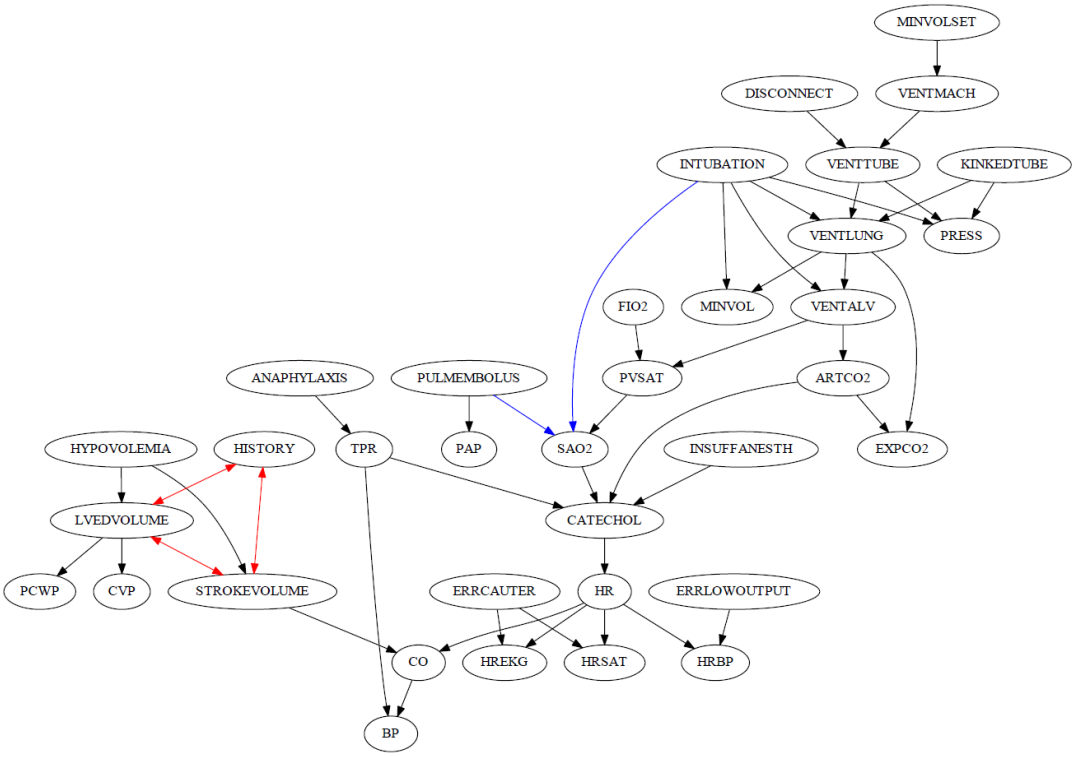

Figure 0.1. The ground truth MAG-5 of the ALARM network with 35 nodes, 46 edges, and latent variables *LVFAILURE* and *SHUNT*. Blue and red edges represent arcs and bi-directed edges in MAG, respectively, that are not present in the ground truth DAG.

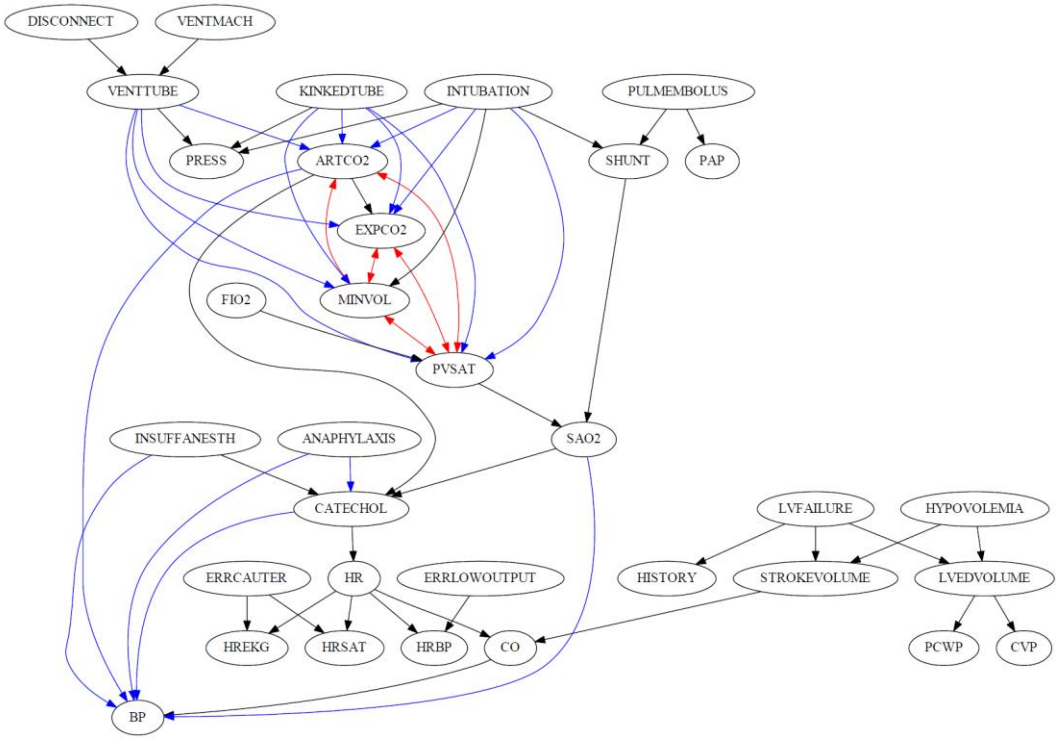

Figure 0.2. The ground truth MAG-10 of the ALARM network with 33 nodes, 55 edges, and latent variables *VENTLUNG*, *MINVOLSET*, *VENTALV*, and *TPR*. Blue and red edges represent arcs and bi-directed edges in MAG, respectively, that are not present in the ground truth DAG.

3.1.2. ASIA network (clean and noisy data)

Table 0.2. The clean and noisy synthetic datasets used for the ASIA network. Each dataset consists of five different sample sizes (from 10^2 to 10^6 samples). For details refer to [23].

Experiment no.	Experiment	Dataset (CSV)	True graph (CSV)	Notes
1	N	Link	DAG (Fig 3)	No noise
2	M5	Link	DAG (Fig 3)	Missing data (5%)
3	M10	Link	DAG (Fig 3)	Missing data (10%)
4	I5	Link	DAG (Fig 3)	Incorrect data (5%)
5	I10	Link	DAG (Fig 3)	Incorrect data (10%)
6	S5	-	-	Merged states data (5%)
7	S10	-	-	Merged states data (10%)
8	L5	-	-	Latent confounders (5%)
9	L10	Link	MAG-10 (Fig 17)	Latent confounders (10%)
10	cMI	Link	DAG (Fig 3)	M5 and I5
11	cMS	-	-	M5 and S5
12	cML	Link	MAG-10 (Fig 17)	M5 and L5
13	cIS	-	-	I5 and S5
14	cIL	Link	MAG-10 (Fig 17)	I5 and L5
15	cSL	-	-	S5 and L5
16	cMISL	Link	MAG-10 (Fig 17)	M5, I5, S5, and L5

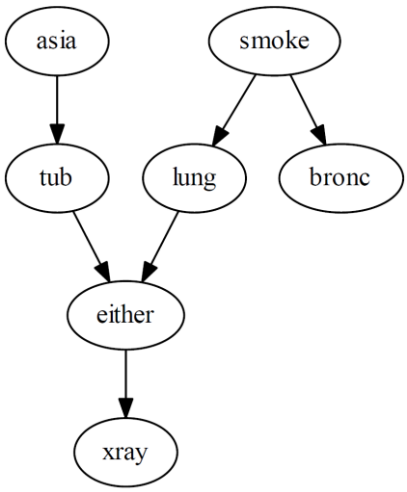

Figure 0.3. The ground truth MAG-10 of the ASIA network with 7 nodes, 6 directed edges, and latent variable *dysp*.

3.1.3. FORMED network (clean and noisy data)

Table 0.3. The clean and noisy synthetic datasets used for the FORMED network. Each dataset consists of five different sample sizes (from 10^2 to 10^6 samples). For details refer to [23].

Experiment no.	Experiment	Dataset (CSV)	True graph (CSV)	Notes
1	N	Link	DAG (Fig 7)	No noise
2	M5	Link	DAG (Fig 7)	Missing data (5%)
3	M10	Link	DAG (Fig 7)	Missing data (10%)
4	I5	Link	DAG (Fig 7)	Incorrect data (5%)
5	I10	Link	DAG (Fig 7)	Incorrect data (10%)
6	S5	Link	DAG (Fig 7)	Merged states data (5%)
7	S10	Link	DAG (Fig 7)	Merged states data (10%)
8	L5	Link	MAG-5 (Fig 18)	Latent confounders (5%)
9	L10	Link	MAG-10 (Fig 19)	Latent confounders (10%)
10	cMI	Link	DAG (Fig 7)	M5 and I5
11	cMS	Link	DAG (Fig 7)	M5 and S5
12	cML	Link	MAG-5 (Fig 18)	M5 and L5
13	cIS	Link	DAG (Fig 7)	I5 and S5
14	cIL	Link	MAG-5 (Fig 18)	I5 and L5
15	cSL	Link	MAG-5 (Fig 18)	S5 and L5
16	cMISL	Link	MAG-5 (Fig 18)	M5, I5, S5, and L5

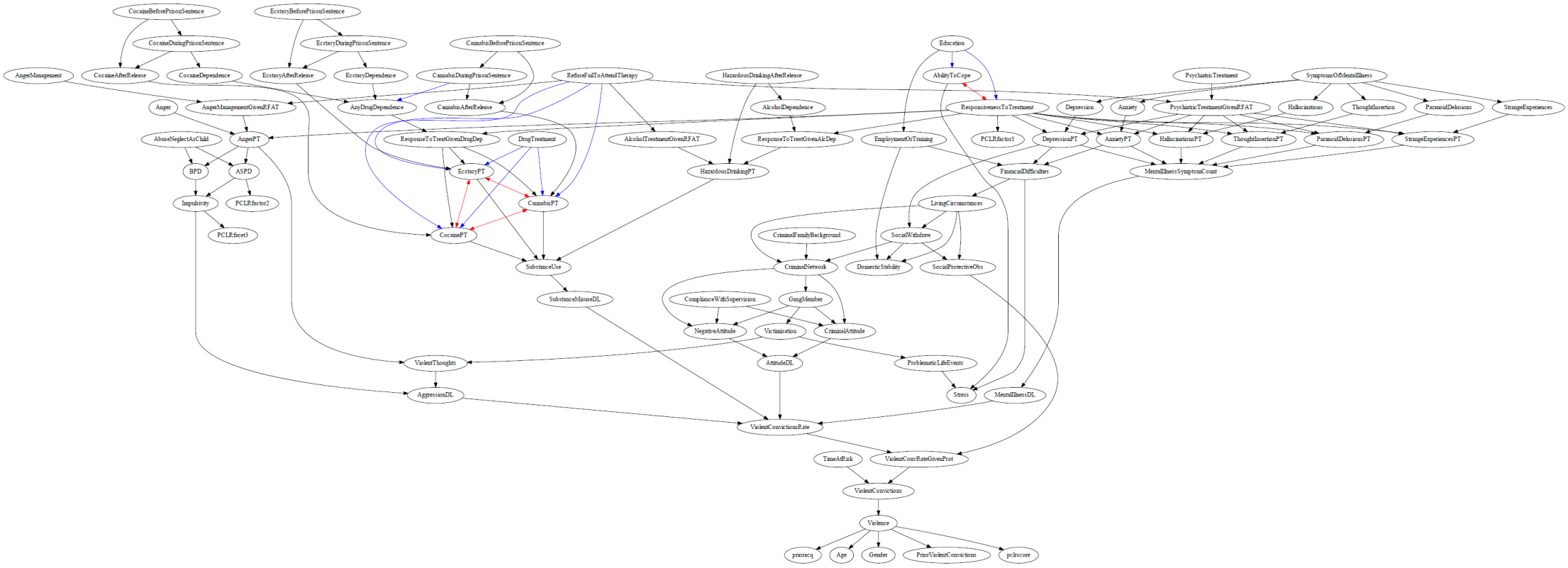

Figure 0.4. The ground truth MAG-5 of the FORMED network with 84 nodes, 140 edges, and *Intelligence*, *CannabisDependence*, *AlcoholTreatment*, and *DrugTreatmentGivenRFAT* as the latent variables. Blue and red edges represent arcs and bi-directed edges in MAG, respectively, that are not present in the ground truth DAG.

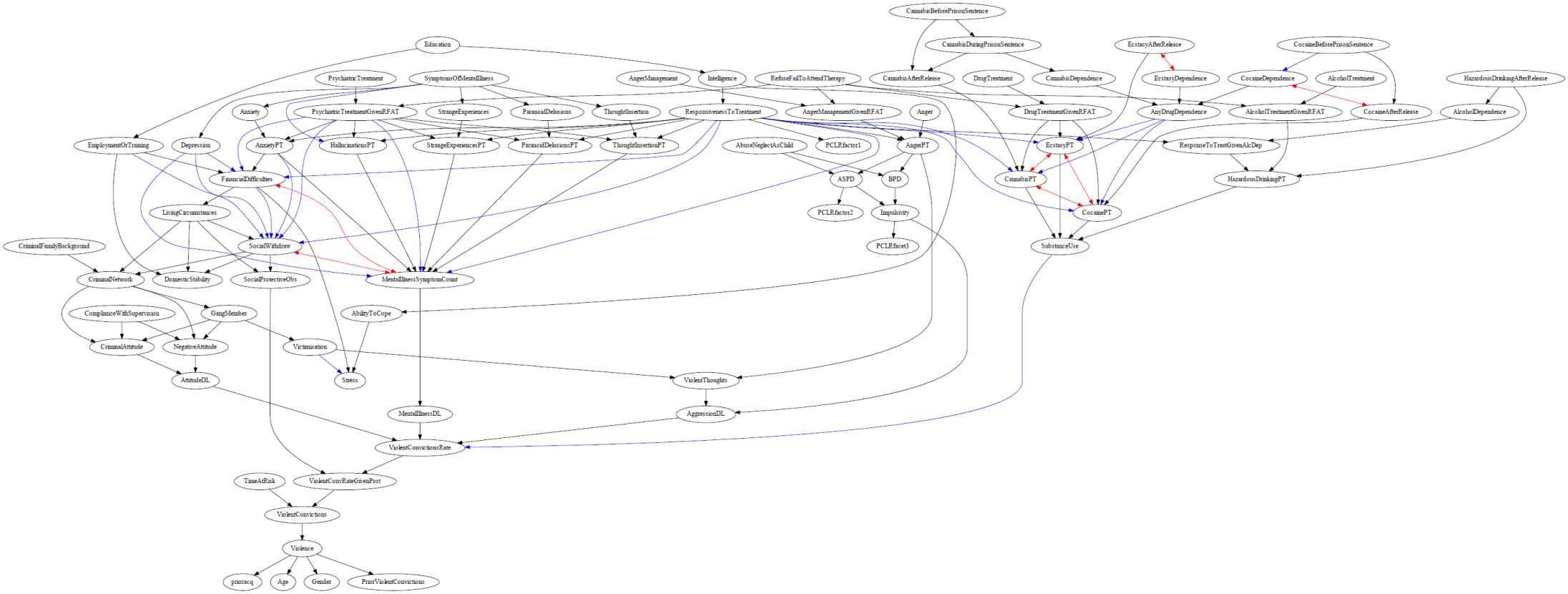

Figure 0.5. The ground truth MAG-10 of the FORMED network with 79 nodes, 142 edges, and *ProblematicLifeEvents*, *pclrscore*, *EcstasyBeforePrisonSentence*, *EcstasyDuringPrisonSentence*, *CocaineDuringPrisonSentence*, *SubstanceMisuseDL*, *DepressionPT*, *ResponseToTreatGivenDrugDep*, and *Hallucinations* as the latent variables. Blue and red edges represent arcs and bi-directed edges in MAG, respectively, that are not present in the ground truth DAG.

3.1.4. PATHFINDER network (clean and noisy data)

Table 0.4. The clean and noisy synthetic datasets used for the PATHFINDER network. Each dataset consists of five different sample sizes (from 10^2 to 10^6 samples). For details refer to [23].

Experiment no.	Experiment	Dataset (CSV)	True graph (CSV)	Notes
1	N	Link	DAG (Fig 9)	No noise
2	M5	Link	DAG (Fig 9)	Missing data (5%)
3	M10	Link	DAG (Fig 9)	Missing data (10%)
4	I5	Link	DAG (Fig 9)	Incorrect data (5%)
5	I10	Link	DAG (Fig 9)	Incorrect data (10%)
6	S5	Link	DAG (Fig 9)	Merged states data (5%)
7	S10	Link	DAG (Fig 9)	Merged states data (10%)
8	L5	Link	MAG-5 (Fig 20)	Latent confounders (5%)
9	L10	Link	MAG-10 (Fig 21)	Latent confounders (10%)
10	cMI	Link	DAG (Fig 9)	M5 and I5
11	cMS	Link	DAG (Fig 9)	M5 and S5
12	cML	Link	MAG-5 (Fig 20)	M5 and L5
13	cIS	Link	DAG (Fig 9)	I5 and S5
14	cIL	Link	MAG-5 (Fig 20)	I5 and L5
15	cSL	Link	MAG-5 (Fig 20)	S5 and L5
16	cMISL	Link	MAG-5 (Fig 20)	M5, I5, S5, and L5

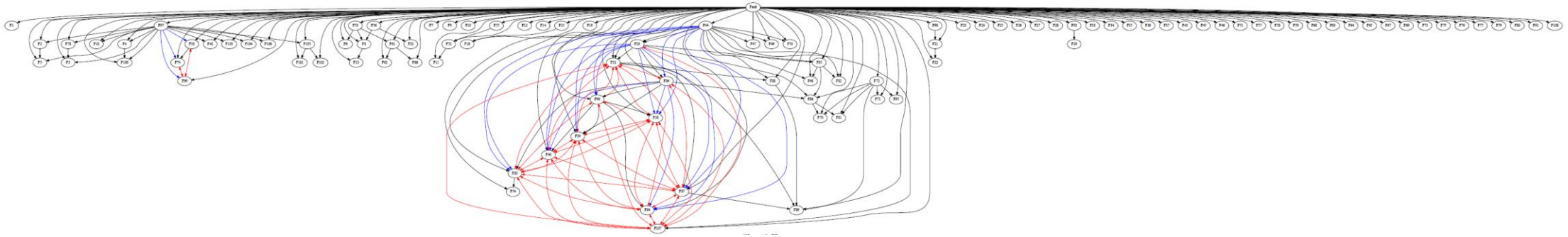

Figure 0.6. The ground truth MAG-5 of the PATHFINDER network with 104 nodes, 230 edges, and latent variables *F17*, *F41*, *F98*, *F68*, *F93*. Blue and red edges represent arcs and bi-directed edges in MAG, respectively, that are not present in the ground truth DAG.

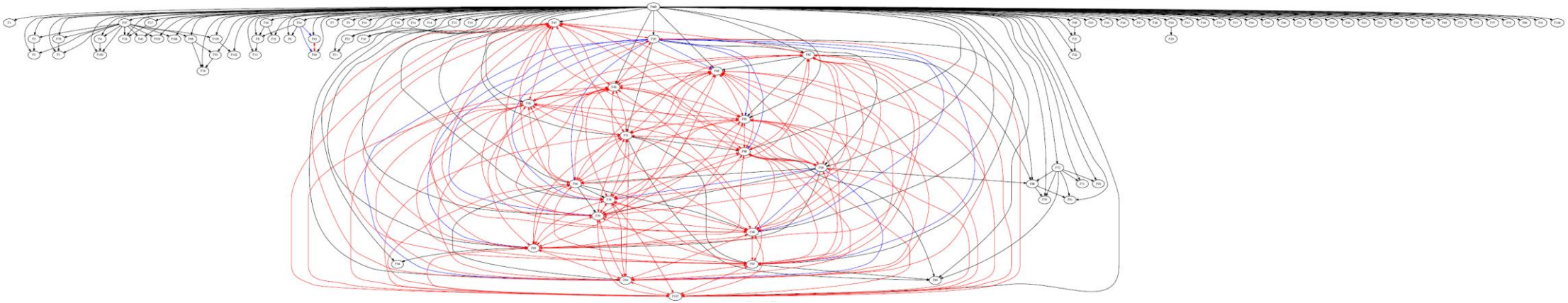

Figure 0.7. The ground truth MAG-10 of the PATHFINDER network with 99 nodes, 297 edges, and latent variables *F41*, *F44*, *F24*, *F36*, *F58*, *F61*, *F76*, *F93*, *F99*, *F101*, *F104*. Blue and red edges represent arcs and bi-directed edges in MAG, respectively, that are not present in the ground truth DAG.

3.1.5. PROPERTY network (clean and noisy data)

Table 0.5. The clean and noisy synthetic datasets used for the PROPERTY network. Each dataset consists of five different sample sizes (from 10^2 to 10^6 samples). For details refer to [23].

Experiment no.	Experiment	Dataset (CSV)	True graph (CSV)	Notes
1	N	Link	DAG (Fig 11)	No noise
2	M5	Link	DAG (Fig 11)	Missing data (5%)
3	M10	Link	DAG (Fig 11)	Missing data (10%)
4	I5	Link	DAG (Fig 11)	Incorrect data (5%)
5	I10	Link	DAG (Fig 11)	Incorrect data (10%)
6	S5	Link	DAG (Fig 11)	Merged states data (5%)
7	S10	Link	DAG (Fig 11)	Merged states data (10%)
8	L5	Link	MAG-5 (Fig 22)	Latent confounders (5%)
9	L10	Link	MAG-10 (Fig 23)	Latent confounders (10%)
10	cMI	Link	DAG (Fig 11)	M5 and I5
11	cMS	Link	DAG (Fig 11)	M5 and S5
12	cML	Link	MAG-5 (Fig 22)	M5 and L5
13	cIS	Link	DAG (Fig 11)	I5 and S5
14	cIL	Link	MAG-5 (Fig 22)	I5 and L5
15	cSL	Link	MAG-5 (Fig 22)	S5 and L5
16	cMISL	Link	MAG-5 (Fig 22)	M5, I5, S5, and L5

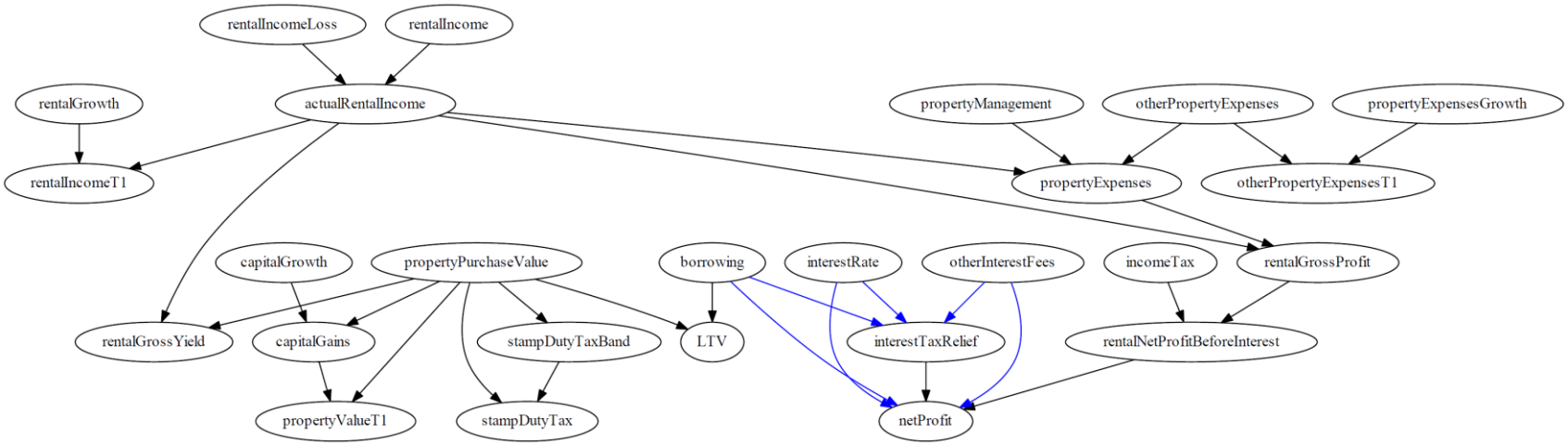

Figure 0.8. The ground truth MAG-5 of the PROPERTY network with 26 nodes, 32 edges, and latent variable *Interest*. Blue edges represent arcs in MAG that are not present in the ground truth DAG.

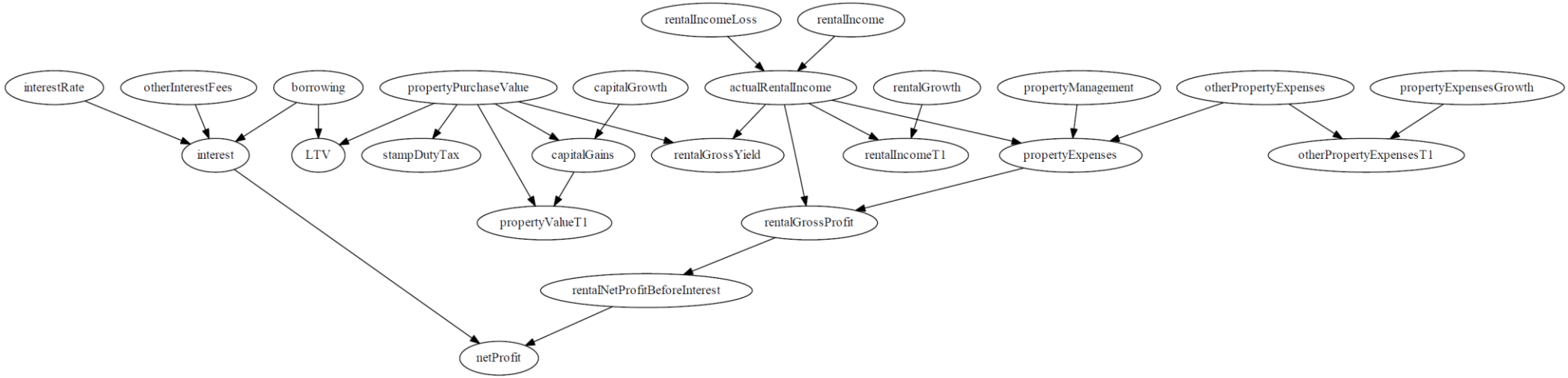

Figure 0.9. The ground truth MAG-10 of the PROPERTY network with 24 nodes, 26 edges, and latent variables *stampDutyTaxBand*, *incomeTax*, and *interestTaxRelief*.

3.1.6. SPORTS network (clean and noisy data)

Table 0.6. The clean and noisy synthetic datasets used for the SPORTS network. Each dataset consists of five different sample sizes (from 10^2 to 10^6 samples). For details refer to [23].

Experiment no.	Experiment	Dataset (CSV)	True graph (CSV)	Notes
1	N	Link	DAG (Fig 13)	No noise
2	M5	Link	DAG (Fig 13)	Missing data (5%)
3	M10	Link	DAG (Fig 13)	Missing data (10%)
4	I5	Link	DAG (Fig 13)	Incorrect data (5%)
5	I10	Link	DAG (Fig 13)	Incorrect data (10%)
6	S5	-	-	Merged states data (5%)
7	S10	Link	DAG (Fig 13)	Merged states data (10%)
8	L5	-	-	Latent confounders (5%)
9	L10	Link	MAG-10 (Fig 24)	Latent confounders (10%)
10	cMI	Link	DAG (Fig 13)	M5 and I5
11	cMS	Link	DAG (Fig 13)	M5 and S5
12	cML	Link	MAG-10 (Fig 24)	M5 and L5
13	cIS	Link	DAG (Fig 13)	I5 and S5
14	cIL	Link	MAG-10 (Fig 24)	I5 and L5
15	cSL	Link	MAG-10 (Fig 24)	S5 and L5
16	cMISL	Link	MAG-10 (Fig 24)	M5, I5, S5, and L5

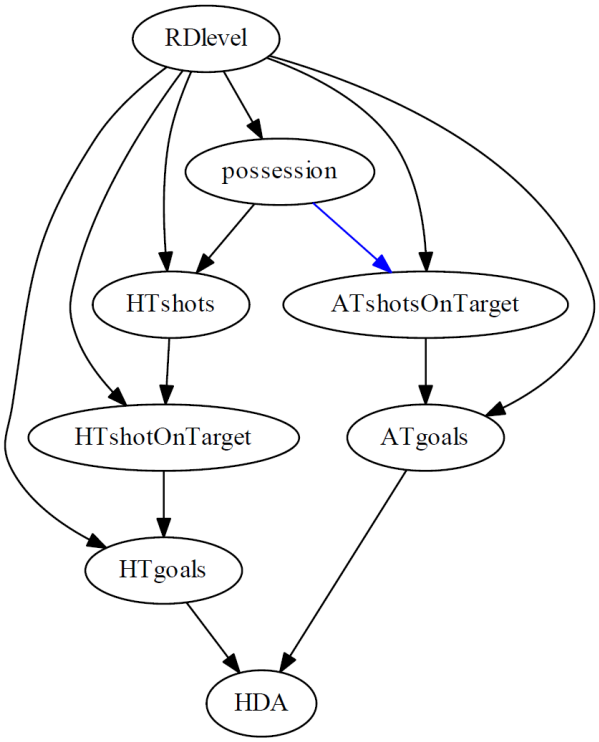

Figure 0.10. The ground truth MAG-10 of the SPORTS network with 8 nodes, 13 edges, and latent variable *ATshots*. Blue edges represent arcs in MAG that are not present in the ground truth DAG.

3.2. Open problems in causal structure learning: A case study of COVID-19 in the UK (2023)

The files available for download below are provided as reference to the work described in [19].

- Download the real data set in all seven different data formats.
- Download the knowledge-based graph which served as the hypothetical ground truth.
- Download all the graphs learnt by the different algorithms.
- Download the seven model-averaging graphs.
- Download all the GeNIe BN models used for cross-validation, interventional and sensitivity analysis.
- Download the source code, input files or software instructions used to test the 29 structure learning algorithms.

3.3. Using GPT-4 to guide causal machine learning (2024)

The files available for download below are provided as reference to the work described in [25].

- Download the four real datasets.
- Download the five graphs constructed by the domain experts.
- Download the GPT-4 prompts; x10 for each of the five case studies.
- Download the GPT-4 outputs; x10 for each of the five case studies.
- Download the GPT-4 averaged outputs; x1 for each of the five case studies.
- Download the GPT-4 constraints used to guide causal structure learning; three different rates of constraints per case study.
- Download the questionnaire responses; 32 participants.

References

[1] Constantinou, A. C. (2018). Bayesian Artificial Intelligence for Decision Making under Uncertainty. *Engineering and Physical Sciences Research Council (EPSRC)*, EP/S001646/1. [Report]

[2] Constantinou, A. C. (2020). Learning Bayesian Networks that enable full propagation of evidence. *IEEE Access*, vol. 8, pp. 124845–123856. [Open-Access DOI]

[3] Constantinou, A. C. (2020). Learning Bayesian Networks with the Saiyan Algorithm. *ACM Transactions on Knowledge Discovery from Data*, vol. 14, Iss. 4, Article 44. [DOI]

[4] Constantinou, A. (2019). The Bayesys user manual. Bayesian AI research lab, MInDS research group, Queen Mary University of London, London, UK. [Online]. Available: http://bayesian-ai.eecs.qmul.ac.uk/bayesys/

[5] Chobtham, K. and Constantinou, A. C. (2020). Bayesian network structure learning with causal effects in the presence of latent variables. In *Proceedings of the 10th International Conference on Probabilistic Graphical Models (PGM-2020)*, Aalborg, Denmark. [PMLR Proceedings download]

[6] Liu, Y., Constantinou, A., and Guo, Z. (2022). Improving Bayesian network structure learning in the presence of measurement error. *Journal of Machine Learning Research*, Vol. 23, Iss. 324, pp. 1–28. [Open-Access DOI]

[7] Liu, Y., and Constantinou, A. (2022). Greedy structure learning from data that contain systematic missing values. *Machine Learning*, Vol. 111, pp. 3867–3896. [Open-Access DOI]

[8] Chobtham, K., Constantinou, A., and Kitson, N. K. (2022). Hybrid Bayesian network discovery with latent variables by scoring multiple interventions. *Data Mining and Knowledge Discovery*, Vol. 37, pp. 476-520 [Open-Access DOI]

[9] Constantinou, A., Liu, Y., Kitson, N. K., Chobtham, K., and Guo, Z. (2022). Effective and efficient structure learning with pruning and model averaging strategies. *International Journal of Approximate Reasoning*, Vol. 151, pp. 292–321. [Open-Access DOI]

[10] Guo, Z. and Constantinou, A. C. (2022). Parallel Sampling for efficient high-dimensional Bayesian network structure learning. arXiv:2202.09691 [cs.LG]

[11] Chobtham, K., and Constantinou, A. (2022). Discovery and density estimation of latent confounders in Bayesian networks with evidence lower bound. In *Proceedings of the 11th International Conference on Probabilistic Graphical Models (PGM-2022)*, Almeria, Spain, Oct 2022. [PMLR Proceedings download]

[12] Beinlich, I. A., Suermondt, H. J., Chavez, R. M., and Cooper, G. F. (1989). The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks. In *Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine*, pp. 247–256.

[13] Lauritzen, S., and Spiegelhalter, D. (1988). Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 50, Iss., pp. 157–224.

[14] Kitson, N. K., & Constantinou, A. (2021). Learning Bayesian networks from demographic and health survey data. Journal of Biomedical Informatics, Vol. 113, Article 103588. [Open-access DOI]

[15] Constantinou, A. C., Freestone, M., Marsh, W., Fenton, N., and Coid, J. (2015). Risk assessment and risk management of violent reoffending among prisoners. *Expert Systems with Applications*, Vol. 42, Iss. 21, pp. 7511–7529. [DOI]

[16] Heckerman, D., Horwitz, E., Nathwani, B. (1992). Towards Normative Expert Systems: Part I. The Pathfinder Project. *Methods of Information in Medicine*, Vol. 31, pp. 90–105.

[17] Constantinou, A. C., and Fenton, N. (2017) The future of the London Buy-To-Let property market: Simulation with Temporal Bayesian Networks. *PLoS ONE*, Vol. 12, Iss. 6, e0179297. [Open Access DOI]

[18] Constantinou, A. (2022). Investigating the efficiency of the Asian handicap football betting market with ratings and Bayesian networks. *Journal of Sports Analytics*, Vol. 8, pp. 171–193. [Open-access DOI]

[19] Constantinou, A., Kitson N. K., Liu, Y., Chobtham, K., Hashemzadeh, A., Nanavati, P. A., Mbuvha, R., and Petrungaro, B. (2023). Open problems in causal structure learning: A case study of COVID-19 in the UK. *Expert Systems with Applications*, Vol. 234, Article 121069 [Open-access DOI]

[20] Zahoor, S., Constantinou, A., Curtis, T. M., and Hasanuzzaman, M. (2024). Investigating the validity of structure learning algorithms in identifying risk factors for intervention in patients with diabetes. arXiv:2403.14327 [cs.LG]

[21] Zahoor, S., Constantinou, A. O'Halloran, F., O'Mahony, L., O'Riordan, M., Kgosidialwa, O., Culliney, L., Alhajri, M. S., and Hasanuzzaman, M. (2025). Causal Insights into Gestational Diabetes Mellitus. TBC.

[22] Constantinou, A., Higgins, N., and Kitson, N. K. (2025). Decoding the mechanisms of the Hattrick football manager game using Bayesian network structure learning for optimal decision-making. arXiv:2504.09499 [cs.LG]

[23] Constantinou, A. C., Liu, Y., Chobtham, K., Guo, Z., and Kitson, N. K. (2021). Large-scale empirical validation of Bayesian Network structure learning algorithms with noisy data. *International Journal of Approximate Reasoning*, vol. 131, pp. 151–188. [Open-Access DOI]

[24] Liu, Y., and Constantinou, A. (2023). Improving the imputation of missing data with Markov Blanket discovery. In *Proceedings of the 11th International Conference on Learning Representations (ICLR-2023)*, Kigali, Rwanda. [Proceedings download]

[25] Constantinou, A., Kitson, N. K, and Zanga, A. (2024). Using GPT-4 to guide causal machine learning. arXiv:2407.18607 [cs.AI]