Appendix A. Background

Causal Graphical Model (CGM) A directed acyclic graph (DAG) is a type of graph G in which the edges e are directed ($\rightarrow$) and there are no cycles. A Causal Graphical Model (CGM) consists of a DAG G and a joint distribution P over a set of random variables $X = (X_1, X_2, \ldots, X_d)$ where P is Markovian with respect to G (Fang and He (2020)). In a CGM, the nodes represent variables X, and the arrows represent causal relationships between them. The joint distribution P can be factorized as follows where $pa(x_i, G)$ denotes the parents of x_i in G.

$$P(x_1, \ldots, x_d) = \prod_{i=1}^{n} P(x_i | pa(x_i, G) \tag{1}$$

A set of DAGs having the same conditional independencies belong to the same equivalence class. DAGs can come in a variety of forms based on the kinds of edges they contain. A Partially Directed Graph (PDAG) contains both directed and undirected edges. A Completed PDAG (CPDAG) consists of directed edges that exist in every DAG G belonging to the same equivalence class and undirected edges that are reversible in G.

Score-based Causal Discovery A score-based causal discovery approach typically searches over the equivalence classes of DAGs to learn the causal graph G that best fits the observed data D as per a score function $S(G, D)$ which returns the score S of G given data D (Chickering (2002), Chowdhury et al. (2023)). Here, the optimization problem for structure learning is as follows:

$$\min_{\substack{G \\ \text{subject to } G \in D}} \quad S(G, X) \tag{2}$$

Typically, any score-based approach has two main components: *(i) a search strategy* - to traverse the search space of candidate graphs G*, and (ii) a score function* - to evaluate the candidate causal graphs.

Score Function A scoring function $S(G, D)$ maps causal DAGs G to a numerical score, based on how well G fits to a given dataset D. A commonly used scoring function to select causal models is the Bayesian Information Criterion (BIC) (Schwarz (1978)) which is defined below:

$$S_{BIC} = -2 * loglikelihood + k * log(n), \tag{3}$$

where n is the sample size used for training and k is the total number of parameters.

Appendix B. Additional Simulation Results

Experimental results of varying the knowledge proportion We present the details of all the metric values for the experiment done with varying the amount of prior knowledge in Table 5. This experiment is done by varying the amount of constraints (directed edges) from 0 to 25 percent each time by raising the amount of knowledge by 5%. The results show that any amount of knowledge is good for improving search accuracy and hence should be leveraged during the search process. Although it is surprising that the increment in knowledge is not directly proportional to the increment in discovery accuracy. Still leveraging any percentage of knowledge is better than using no knowledge at all.

Table 5: Results (metric values) of the experiment done by varying knowledge proportion.

Datasets	Proportion of Knowledge	SHD (lower better)	TPR (higher better)	FDR (lower better)	Estimated models (lower better)
Child	0%	34	0.38	0.89	49
	5%	26	0.62	0.79	39
	10%	**21**	**0.69**	**0.74**	36
	15%	26	**0.69**	0.77	37
	20%	24	**0.69**	0.76	35
	25%	24	0.62	0.78	**34**
Alarm	0%	56	0.74	0.61	84
	5%	56	0.74	0.61	82
	10%	54	0.79	0.59	83
	15%	53	0.79	0.58	76
	20%	53	0.81	0.59	77
	25%	**51**	**0.84**	**0.57**	**74**
Hepar2	0%	70	0.5	0.23	83
	5%	59	0.58	**0.15**	81
	10%	67	0.54	0.21	76
	15%	64	0.55	0.21	73
	20%	**55**	**0.64**	0.17	71
	25%	58	0.6	0.21	**66**

Appendix C. Baseline Causal Discovery Approaches

We report the performance of different baseline causal discovery approaches such as PC (constraint-based), GES (score-based) and LiNGAM (FCM-based) on the experimental datasets to see their comparative performance with respect to KGS. We briefly discuss the methods below:

(i) PC: The Peter-Clark (PC) algorithm (Spirtes et al. (2000)) is a very common constraint-based causal discovery approach that largely depends on conditional independence (CI) tests to find the underlying causal graph. Primarily, it works in three steps: (i) Skeleton construction, (ii) V-structures determination, and (iii) Edge orientations.

(ii) GES: Greedy Equivalence Search, GES (Chickering (2002)) is one of the oldest score-based causal discovery methods that employ a greedy search over the space of equivalence classes of DAGs. Primarily GES operates in two phases: (i) Forward Equivalence Search (FES) and (ii) Backward Equivalence Search (BES). GES assumes a decomposable score function $S(G, D)$ which is expressed as a sum of the scores of individual nodes and their parents. A problem with GES is that the number of search states that it needs to evaluate scales exponentially with the number of nodes d in the graph (Chickering and Meek (2015)). This results in a vast search space, and also scoring a large number of graphs which adds to the overall cost as score computation is an expensive step.

$$S(G, D) = \sum_{i=1}^{d} s(x_i, pa(x_i, G)) \tag{4}$$

(iii) LiNGAM: Linear Non-Gaussian Acyclic Model (LiNGAM) uses a statistical method known as independent component analysis (ICA) to discover the causal structure from observational data. It makes some strong assumptions such as the data generating process is linear, there are no unobserved confounders, and noises have non-Gaussian distributions with non-zero variances (Shimizu et al. (2006)).

DirectLiNGAM (dLiNGAM) is an efficient variant of the LiNGAM approach that uses a direct method for learning a linear non-Gaussian structural equation model (Shimizu et al. (2011)). The direct method estimates causal ordering and connection strengths based on non-Gaussianity.

Appendix D. Performance Metrics

• **Structural Hamming Distance (SHD):** SHD is the sum of the edge additions (A), deletions (D) or reversals (R) that are required to convert the estimated graph into the true causal graph (Zheng et al. (2018); Cheng et al. (2022)). To estimate SHD it is required to determine the missing edges, extra edges and edges with wrong direction in the estimated graph compared to the true graph. Lower the SHD closer is the graph to the true graph and vice versa. The formula to calculate SHD is given below:

$$SHD = A + D + R \tag{5}$$

• **True Positive Rate (TPR):** TPR denotes the proportion of the true edges in the actual graph that are correctly identified as true in the estimated graph. A higher value of the TPR metric means a better causal discovery.

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

Here, TP means the true positives or the number of correctly identified edges and FN or false negatives denote the number of unidentified causal edges.

• **False Discovery Rate (FDR):** FDR is the ratio of false discoveries among all discoveries (Zheng et al. (2018)). FDR represents the fraction of the false edges over the sum of the true and false edges. Lower the FDR, better is the outcome of causal discovery.

$$FDR = \frac{FP}{TP + FP} \tag{7}$$

Here, FP or false positives represent the number of wrongly identified directed edges.

Appendix E. Extraction of Causal Priors using LLMs

List of the academic papers relevant to oxygen therapy dataset used by GPT-4:

1. Does age matter? The relationship between age and mortality in penetrating trauma (Ottochian et al. (2009)).

Table 6: The causal relationships retrieved by GPT-4 from relevant literature papers and their corresponding prompt-answer link.

Edge by GPT-4	**Chat Link of GPT-4 Prompts and Answers**
PaO2 → SpO2	https://chat.openai.com/share/5dc1db93-b394-4bcd-a5de-837d0863e31f
SpO2→PaO2	https://chat.openai.com/share/cff9f68c-4bb7-4f23-86a8-056aa7fd1955
FiO2→SpO2	https://chat.openai.com/share/1b3a5e3c-d84d-41d7-a2c4-263ebb6766f6
PaO2 → SOFA	https://chat.openai.com/share/07fa3389-e8ec-41f2-b45e-71c0496c7852
ARDS→SpO2	https://chat.openai.com/share/31644194-7aed-4ad5-95ae-bcfc4c8ec6d1
PEEP→PaO2	https://chat.openai.com/share/7dccd77e-3e0d-4d05-9624-2957148c4490
COPD→SpO2	https://chat.openai.com/share/6fb4857c-1656-402e-8eae-d56f6c2f8c20
COPD→FiO2	https://chat.openai.com/share/b576b2eb-4d73-484e-9cac-99fc02bcbe92
COPD→PaCO2	https://chat.openai.com/share/32c83464-9498-42b6-bd15-2f94fb2d0d9c
FiO2→PaO2	https://chat.openai.com/share/66c04507-a981-42a8-8d89-f56eff9521c4
Age→Trauma	https://chat.openai.com/share/a49042e8-ea48-42a6-a0c2-2ae1b4b9492e
Age→Death	https://chat.openai.com/share/daca5caf-48b7-4168-988a-7ae3e44ccaf8
SOFA→SaO2	https://chat.openai.com/share/43659587-c1da-4714-a262-1194671cc9c5
SpO2→Death	https://chat.openai.com/share/ee3ce03a-7951-4999-9dd2-395665871838
VT→Oxygenation	https://chat.openai.com/share/57905bad-9972-4e37-8f1b-912127ebd226
Oxygenation→Death	https://chat.openai.com/share/79134416-c2fa-476b-b790-bf9e8bab7d89
FiO2→SaO2	https://chat.openai.com/share/cc09e77b-589e-4ee9-a4db-5add3d81b69a
COPD→PaCO2	https://chat.openai.com/share/919b0fa4-ea06-4097-8578-058303852941

2. Conservative versus Liberal Oxygenation Targets for Mechanically Ventilated Patients A Pilot Multicenter Randomized Controlled Trial (Panwar et al. (2016)).

3. Conservative Oxygen Therapy during Mechanical Ventilation in the ICU (Investigators et al. (2020)).

4. Effect of Conservative vs Conventional Oxygen Therapy on Mortality Among Patients in an Intensive Care Unit The Oxygen-ICU Randomized Clinical Trial (Girardis et al. (2016)).

5. Liberal or Conservative Oxygen Therapy for Acute Respiratory Distress Syndrome (Barrot et al. (2020)).

6. Effect of High-Flow Nasal Oxygen vs Standard Oxygen on 28-Day Mortality in Immunocompromised Patients With Acute Respiratory Failure The HIGH Randomized Clinical Trial (Azoulay et al. (2018)).

7. Oxygen Therapy in Chronic Obstructive Pulmonary Disease (Kim et al. (2008)).

8. Effect of Low-Normal vs High-Normal Oxygenation Targets on Organ Dysfunction in Critically Ill Patients A Randomized Clinical Trial (Gelissen et al. (2021)).

9. Oxygen-Saturation Targets for Critically Ill Adults Receiving Mechanical Ventilation (Semler et al. (2022)).

10. Mechanical Ventilation: State of the Art (Pham et al. (2017)).

To avoid any mistakes we did not blindly rely on the LLM's statements, rather we verfied those by an expert before incorporating in the ground truth graph.

Appendix F. Ground Truth Graphs

Oxygen Therapy G_T Figure 5 represents the ground truth graph of the OT dataset available in the study Gani et al. (2023). This graph has total 66 causal edges. We used 64 of them exactly, reversed 2 edges and augmented 3 newly obtained edges. The 2 reversed edges are: FiO2→SpO2 and COPD→SpO2. The 3 added edges are ARDS→SpO2, FiO2→PaO2 and FiO2→SaO2. Other than these modifications, all other edges in Figure 5 have been kept unaltered in the ground truth that we used (see Figure 6) for our experiments.

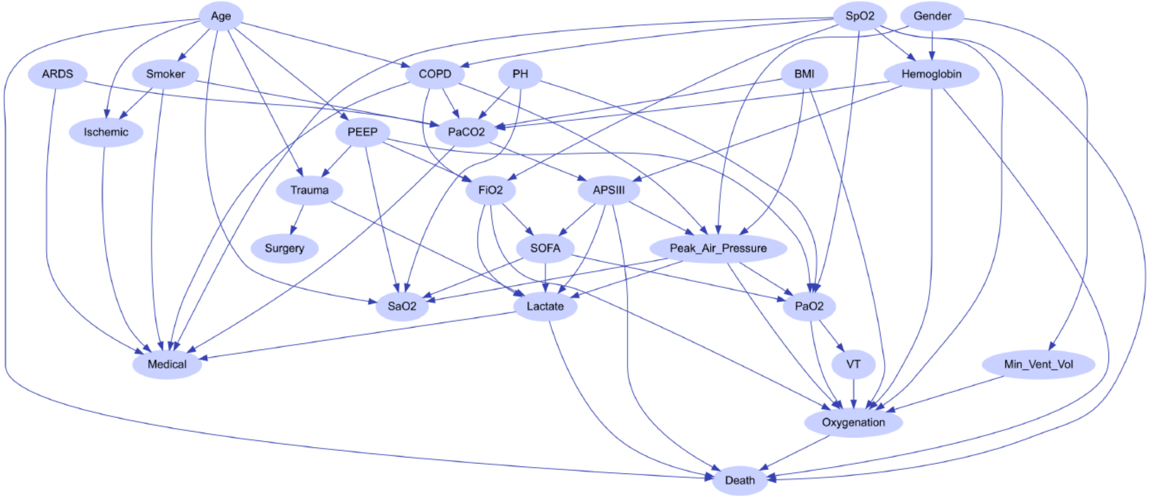

Figure 5: Ground truth graph of Oxygen Therapy dataset from Gani et al. (2023).

Child G_T The ground truth graph of the Child network is available in https://www.bnlearn.com/bnrepository/discrete-medium.html#child.

Alarm G_T The ground truth graph of the Alarm network is available in https://www.bnlearn.com/bnrepository/discrete-medium.html#alarm.

Hepar2 G_T The ground truth graph of the Hepar2 network is available in https://www.bnlearn.com/bnrepository/discrete-large.html#hepar2.

Appendix G. Code Availability

After the blind review period is over, we will add a link to a public repository for the code and datasets. For now, we have uploaded the code and datasets of KGS for review as supplementary material.

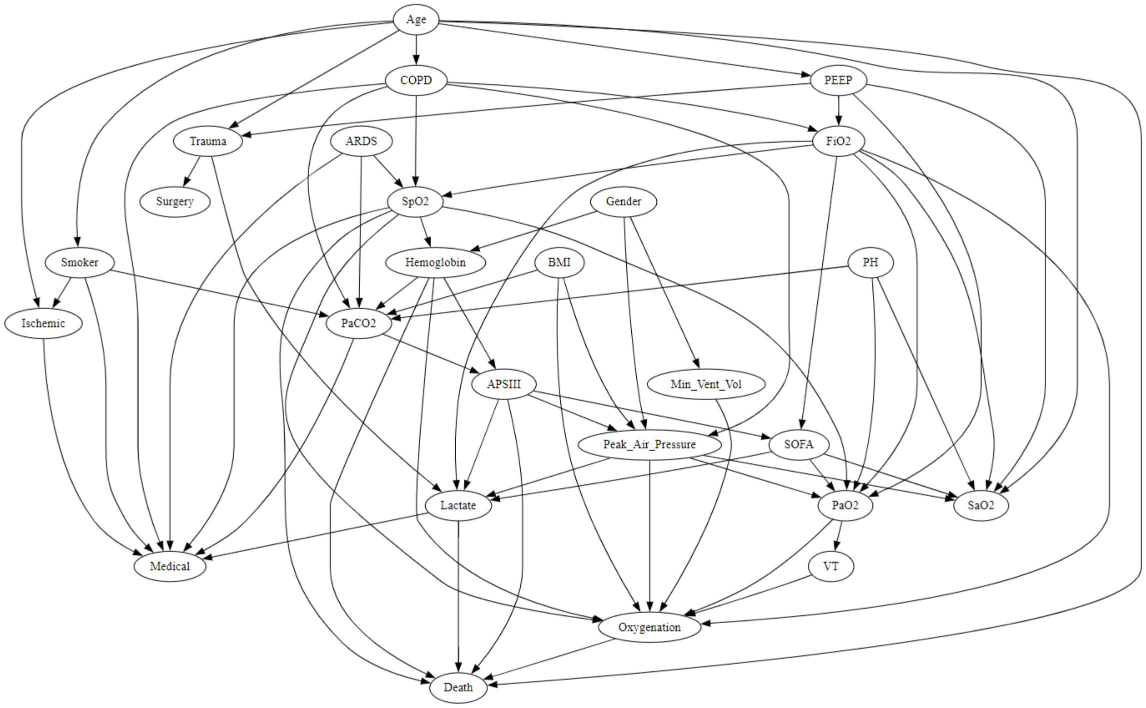

Figure 6: Reformed ground truth graph of Oxygen Therapy dataset with 69 causal edges that we used for our experiments. Here, the reversed edges are: FiO2→SpO2 and COPD→SpO2. The added edges are ARDS→SpO2, FiO2→PaO2 and FiO2→SaO2.