Causal Discovery in Additive Noise Models using Beam Search

Hans Jarett J. Ong[1], Brian Godwin S. Lim[2,3], Renzo Roel P. Tan[4,5], Kazushi Ikeda[6]

[1]Division of Information Science, Nara Institute of Science and Technology, Nara, Japan
(E-mail: ong.hans_jarett.ol5@naist.ac.jp)
[2]Division of Information Science, Nara Institute of Science and Technology, Nara, Japan
(E-mail: brian.lim@naist.ac.jp)
[3]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[4]Division of Information Science, Nara Institute of Science and Technology, Nara, Japan
(E-mail: rr.tan@is.naist.jp)
[5]John Gokongwei School of Management, Ateneo de Manila University, Metro Manila, Philippines
(E-mail: rrtan@ateneo.edu)
[6]Division of Information Science, Nara Institute of Science and Technology, Nara, Japan
(E-mail: kazushi@is.naist.jp)

Abstract: Causal discovery from observational data is a fundamental challenge. Greedy search algorithms like Regression with Subsequent Independence Test (RESIT), commonly used for learning Additive Noise Models (ANMs), are susceptible to making irreversible errors, especially in high-variance contexts. Such settings can be caused by unmeasured confounders or by high statistical noise from finite samples. To address this, we introduce a novel generalization of RESIT that replaces its local, greedy search with a more robust beam search, framing the task as a path search on a state-space graph. Through extensive simulation experiments, we demonstrate that structural accuracy, measured by Structural Hamming Distance (SHD) and Structural Intervention Distance (SID), consistently improves as the beam width (w) increases. Crucially, we also show that this performance gain comes at a manageable, approximately linear increase in computational cost relative to w. Furthermore, our analysis across different sample sizes shows these gains are most statistically significant in intermediate regimes ($n = 250, 500$). This suggests that at these sample sizes, the statistical noise is high enough to mislead the greedy search into a suboptimal ordering, an error our wider beam search corrects, while performance converges at large sample sizes ($n = 1000$). Our framework provides a practical, tunable algorithm that bridges the gap between fast but brittle local search methods and computationally infeasible global searches, thereby enhancing the reliability of causal discovery in complex, high-variance settings where such local errors are common.

Keywords: Causal Discovery, Additive Noise Models, Beam Search, Structural Equation Models, Unmeasured Confounders

1. INTRODUCTION

A fundamental goal in fields like robotics and artificial intelligence is to create autonomous agents that can not only observe patterns but also understand the underlying mechanics of their environment. To act intelligently, an agent must predict the consequences of its actions—a task requiring a shift from correlation to causation [1, 2]. While the gold standard for establishing causality is the randomized controlled trial (RCT) , performing such experiments is often impractical, unethical, or impossible in complex domains like robotics, biology, or economics. This limitation motivates the field of *causal discovery*, which seeks to infer causal structures directly from readily available observational data. Unlike data from an RCT, observational data is collected by passively observing a system in its natural state, without any manipulation from the researcher or agent.

Learning causal structure from such observational data is, however, an ill-posed problem. Without additional assumptions, different causal structures can be statistically indistinguishable, often allowing for the identification of only a set of possible structures, known as a *Markov Equivalence Class* (MEC) [3]. A powerful way to overcome this ambiguity is to impose constraints on the functional form of the causal relationships, leading to model families like the Additive Noise Model (ANM) [4]. The Regression with Subsequent Independence Test (RESIT) algorithm was developed as a practical method for learning ANM structures, employing a stepwise *greedy search* to find the causal ordering of the variables.

However, the greedy nature of RESIT is also its primary weakness. The algorithm is susceptible to local optima: an incorrect decision made early in the search is final and cannot be undone, potentially leading to a cascade of errors. This vulnerability is particularly problematic in practical settings where the locally optimal choice may be incorrect due to multiple factors. A primary reason is the violation of *causal sufficiency*—the assumption that all common causes of the observed variables have been measured [2, 3]. An unobserved common cause (i.e., unmeasured confounder) can induce a spurious correlation between its effects, creating the illusion of a direct causal link. Furthermore, even if causal sufficiency holds, *statistical noise from finite samples* can cause an incorrect variable to be selected as the sink at any given step, leading to propagated errors.

A promising path forward has been proposed in work on the Linear Non-Gaussian Acyclic Model (LiNGAM) fam-

† Hans Jarett J. Ong is the presenter of this paper.

ily, a framework restricted to linear relationships and non-Gaussian noise [5]. Like RESIT, a prominent algorithm in this family, DirectLiNGAM, also employs a greedy search to find the causal ordering [6]. To address the limitations of this local search, several extensions have been proposed. One line of work aims for a global solution: the LiNGAM-MMI method, for example, reformulates the task as a shortest-path problem on a state-space graph [7]. This seeks the optimal causal ordering that minimizes the total statistical dependence. Another line of work has explored heuristic solutions, such as applying beam search to DirectLiNGAM, to find a more robust ordering without the computational cost of a full global search [8].

These extensions, however, have so far been limited to the linear case. Many systems of interest in artificial life and robotics exhibit complex, non-linear dynamics, making an extension to the non-linear domain highly desirable. A direct application of a global search (akin to LiNGAM-MMI) to the non-linear RESIT framework, however, is computationally infeasible due to the factorial growth of the search space. Therefore, we adopt the alternative heuristic approach. We propose to generalize RESIT by employing a **beam search** algorithm, a method that has proven effective in the linear case [8] but has not yet been extended to non-linear ANMs.

This beam search approach provides a unifying framework for these search-based methods. The standard greedy search of RESIT is thus equivalent to a beam search with a width of one ($w = 1$), while an optimal best-first search (i.e., a global search) becomes equivalent to an infinite beam width ($w \to \infty$), positioning the proposed framework as a generalization of these extreme approaches. By selecting a finite beam width, our proposed method becomes more robust than a simple local search while avoiding the prohibitive computational cost of a full global search. This paper makes the following contributions:

- We generalize the greedy search of RESIT by framing causal ordering for ANMs as a path search problem on a state-space graph, a concept previously explored only in the linear case (LiNGAM).
- We propose a practical beam search algorithm to approximate the optimal solution to this problem, offering a tunable trade-off between the local search of RESIT and a computationally infeasible optimal (or global) search.
- We demonstrate through simulation experiments that our beam search approach significantly outperforms the standard greedy RESIT by providing enhanced robustness against local estimation errors caused by statistical noise from finite samples or unmeasured confounders. We show these performance gains are most statistically significant in intermediate, high-variance sample regimes ($n = 250, 500$), where our method successfully navigates noisy local optima that trap a purely greedy search.

2. PRELIMINARIES

2.1. Graphical Causal Models

To formalize causal relationships, we use the framework of Structural Causal Models (SCMs) [2]. An SCM consists of a set of variables $\mathbf{X} = \{X_1, \ldots, X_d\}$, a directed acyclic graph (DAG) $\mathcal{G}$ over these variables, and a set of functions $\{f_1, \ldots, f_d\}$ that assign a value to each variable. The value of each variable X_i is determined by a function f_i of its direct causes (its parents PA_i in $\mathcal{G}$) and an exogenous random noise term N_i:

$$X_i \coloneqq f_i(\text{PA}_i, N_i). \tag{1}$$

The noise terms $N_1, \ldots, N_d$ represent all factors that are not explicitly modeled but affect the variables. A core assumption in many causal discovery methods is that these noise terms are mutually independent, i.e., $N_i \perp\!\!\!\perp N_j$ for all $i \neq j$. This independence assumption implies the *causal sufficiency* assumption (i.e., the absence of unmeasured confounders) mentioned in Section 1. If there were an unmeasured confounder of two variables X_i and X_j, then after accounting for their observed parents, the remaining variations in both variables (represented by N_i and N_j) would still be influenced by this common cause, making them statistically dependent.

2.2. Additive Noise Models (ANMs)

As discussed in Section 1, a central challenge in causal discovery on observational data is that causal DAGs generally can only be determined up to their MEC. To overcome this ambiguity and identify a single DAG, one powerful approach is to assume the data was generated by an SCM with a specific functional form. These assumptions break the statistical symmetries that make different causal directions indistinguishable, allowing for the identification of a unique graph from the MEC.

Additive Noise Models (ANMs) are a class of SCMs that impose such a restriction [4]. In an ANM, the function f_i is constrained so that the noise term is additive:

$$X_i \coloneqq f_i(\text{PA}_i) + N_i. \tag{2}$$

This specific structure is powerful because it breaks the statistical symmetry between cause and effect, allowing the true causal DAG to be identified. While there are several other identifiable model classes (e.g., post-nonlinear models), this work focuses on ANMs. Well-known cases when ANMs are identifiable include:

1. When the functions f_i are non-linear (for any noise distribution, including Gaussian) [4].
2. When the functions f_i are linear, but the noise terms N_i are non-Gaussian [5]. This subclass is often referred to as LiNGAM.

In this paper, we focus primarily on the non-linear case.

2.3. RESIT: A Greedy Search for ANMs

The RESIT algorithm is a practical method for learning ANM structures from data [4]. It works by determining the causal order of the variables iteratively. The core idea is that the noise term N_j must be independent of PA_j, i.e., the parents of X_j. RESIT exploits a direct consequence of this core idea: for any "sink" node X_j (a variable with no effects in the system), its corresponding noise term N_j will be independent of all other variables.

At each step k of the algorithm, RESIT searches for a sink node among the set of remaining candidate variables S_k. It solves a local optimization problem by treating each candidate variable $X_j \in S_k$ as a potential effect and regressing it on all other remaining variables $S_k \setminus \{X_j\}$. The variable $X_{j^*}^{(k)}$ whose residuals $r_{j^*}^{(k)}$ (the value of the noise term estimated from the regression) are most independent of its regressors is chosen as the sink:

$$X_{j^*}^{(k)} = \arg\min_{X_j \in S_k} \text{dep}(r_j^{(k)}, S_k \setminus \{X_j\}), \tag{3}$$

where $\text{dep}(\cdot, \cdot)$ is a measure of statistical dependence. In the original paper, the p-value of an independence test is maximized, which is equivalent to minimizing dependence. This procedure is a *greedy* search because at each step, it commits to the variable that is locally optimal. The choice is final and cannot be revised, even if it leads to a suboptimal global solution. If an incorrect variable is chosen as the sink due to statistical noise from finite samples or the presence of unmeasured confounders (i.e., a violation of causal sufficiency), the error propagates through all subsequent steps, potentially leading to a highly inaccurate final graph.

2.4. From Local to Global Search: The Linear Case

The greedy search strategy employed by RESIT is not unique to non-linear models. In the linear subclass of ANMs, known as LiNGAM [5], a similar greedy algorithm called DirectLiNGAM exists [6]. While both RESIT and DirectLiNGAM are greedy search algorithms, they determine the causal order in opposite directions. RESIT proceeds from *sink-to-source*. At each step, it identifies an effect (a sink) by finding the variable whose noise term is most independent of all other remaining variables. In contrast, DirectLiNGAM proceeds from *source-to-sink*. It identifies a cause (a source) by finding the variable that is most independent of the residuals of all other variables regressed upon it.

At each step k, DirectLiNGAM searches for a source node among the set of remaining candidate variables S_k. For each candidate source $X_j \in S_k$, it computes a corresponding set of residuals, $\mathcal{R}_j^{(k)}$, by regressing every other variable $X_i \in S_k \setminus \{X_j\}$ on X_j. The algorithm then selects the variable, denoted $X_{j^*}^{(k)}$, that is most independent of its corresponding set of residuals, $\mathcal{R}_{j^*}^{(k)}$. To proceed to the next step, the influence of this identified source is removed by defining the set of variables for the next iteration, S_{k+1}, as this new set of residuals, $\mathcal{R}_{j^*}^{(k)}$. The algorithm repeats this procedure until a full causal ordering is established.

However, like RESIT, the fundamental limitation of DirectLiNGAM is its susceptibility to local optima. This shared vulnerability of greedy methods motivates the development of more robust, global search strategies. To address this limitation, the LiNGAM-MMI method was proposed, which generalizes the local (greedy) search of DirectLiNGAM into a global search [7]. This is achieved by reformulating the causal ordering task as a shortest-path finding problem on a state-space graph. In this formulation, a path through the state-space graph corresponds to a full causal ordering, and the total path cost serves as a global measure of the joint independence of the system's noise terms. While this global approach is more robust against local optima, finding the shortest path with an optimal best-first search algorithm like Dijkstra's is often computationally infeasible, as the number of possible orderings grows factorially with the number of variables. This trade-off between search robustness and computational cost is the primary motivation for our proposed method.

3. BEAM SEARCH FOR ANM CAUSAL DISCOVERY

To balance the trade-off between the local sub-optimality of greedy search and the computational cost of a global best-first search, we propose a novel causal discovery method for ANMs based on beam search. Our method consists of two main phases: first, we determine the causal ordering of the variables using our beam search approach; second, we prune the resulting fully-connected graph to identify the final causal DAG using the same procedure as RESIT [4].

3.1. Phase 1: Causal Ordering via Beam Search

3.1.1. Search Framework

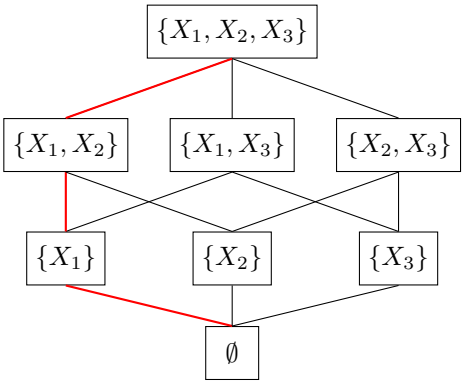

Fig. 1. The state-space graph for causal ordering. Nodes represent the set of variables not yet ordered. An edge from a set S_{k-1} to $S_{k-1} \setminus \{X_j\}$ represents selecting X_j as the sink at step k. The cost is the dependence between the residuals of $X_j \sim (S_{k-1} \setminus \{X_j\})$ and the regressors. The red path shown corresponds to identifying the reverse causal order (X_3, X_2, X_1), which yields the final causal order $X_1 \to X_2 \to X_3$.

Inspired by the global search formulation in the linear case [7], we frame the causal ordering task for ANMs as a search for a low-cost path in the state-space graph illustrated in Fig. 1. The core principle of this search is that the noise (estimated by the residuals) must be independent of the causes. Therefore, the cost for selecting a variable X_j as the sink at step k from a set of candidates S_{k-1} is defined as the statistical dependence between the residuals $r_j^{(k)}$ (from regressing X_j on the remaining variables $S_{k-1} \setminus \{X_j\}$) and its regressors:

$$\text{cost}(S_{k-1} \to S_{k-1} \setminus \{X_j\}) = \text{dep}(r_j^{(k)}, S_{k-1} \setminus \{X_j\}). \tag{4}$$

In our implementation, we use the Hilbert-Schmidt Independence Criterion (HSIC) [9] test statistic as the dependence

measure $\text{dep}(\cdot, \cdot)$, as it provides a robust, kernel-based test for independence. The total cost of a path is the sum of these individual step costs. The goal is to find a path with low total cost. This total cost serves as a global score for an entire ordering, and minimizing it allows the search to escape the local optima limitation of a purely greedy search.

Using the test statistic's magnitude for path scoring is a deliberate choice. In a greedy search ($w = 1$), this is equivalent to maximizing the p-value, as done in the original RESIT paper [4]. However, for a beam search ($w > 1$), this choice is critical. Aggregating path cost by summing p-values is not statistically meaningful. Our approach of summing the test statistic provides a direct heuristic for the "total dependence" of an ordering and is consistent with the implementation of RESIT in the widely-used `lingam` library [10]. We still use p-values in Phase 2 (Section 3.2), as that phase involves hypothesis testing for pruning. Another potential approach for scoring could be to aggregate p-values along a path, for instance using Fisher's method. We leave this as an interesting direction for future exploration, as our chosen method of summing the test statistic aligns directly with our objective of minimizing the overall measured dependence.

3.1.2. Beam Search Algorithm

To find a low-cost path without the computational expense of an optimal best-first search, we employ a beam search. A true optimal best-first search algorithm like Dijkstra's is guaranteed to find the optimal path by keeping all possible partial paths in memory and always expanding the one with the lowest current cost. Beam search approximates this process by following the same principle but with a crucial heuristic: at each step, it prunes the set of candidate paths, keeping only a predetermined number, w (the **beam width**), of the most promising solutions. While this trade-off means the algorithm is no longer guaranteed to find the globally optimal path, it allows for a much more efficient exploration of the search space [11].

The algorithm proceeds through d selection steps to build a complete reverse ordering of the d variables.

1. **Initialization:** Start with a beam containing a single path hypothesis: the initial state (the set of all variables) with a total cost of 0.
2. **Iteration:** For each of the d selection steps:
   - **Expand:** Generate all possible next states from every path currently in the beam. A next state is formed by selecting one variable from a path's current set to be the next sink.
   - **Evaluate:** For each newly generated candidate path, calculate its total cost by adding the new step cost to the accumulated cost of its parent path. The step cost is calculated using the HSIC test statistic, which quantifies the degree of dependence.
   - **Prune:** From the collection of all candidate paths generated in this step, select the top w paths with the lowest total costs. The parameter w is the **beam width**. These w paths form the new beam for the next iteration, while all other candidates are discarded.
3. **Termination:** The search terminates after exactly d steps. Since the task is to produce a complete ordering of d variables, a full solution path must contain exactly d selections. After d iterations, the beam contains a set of complete candidate orderings, from which the one with the lowest total cost is selected as the solution.

The beam width w is a crucial hyperparameter that controls the exploration-exploitation trade-off of the search:

- If $w = 1$, the algorithm reduces to a standard greedy search, i.e., the existing RESIT method, as it only ever considers the single best choice at each step. This serves as our primary baseline in the experiments.
- If $w \to \infty$, the algorithm becomes equivalent to an optimal best-first search, guaranteeing that the true shortest path is found.

By choosing a moderate value for w, we can explore a much larger portion of the search space than a greedy algorithm while remaining computationally feasible.

3.2. Phase 2: Causal DAG Identification by Pruning

Once the optimal causal order is found, the second phase of our method identifies the final causal DAG. This phase is identical to the edge pruning procedure used in the original RESIT algorithm [4]. It begins with a fully connected directed acyclic graph (DAG) that is consistent with the determined causal order.

The core of the pruning logic is an iterative procedure that tests each potential parent edge for removal. For each variable X_i in the determined causal order, we consider its set of candidate parents, PA_i, which includes all of its predecessors in the order. For each candidate parent $X_j \in \text{PA}_i$, we perform a specific independence test. First, we compute the residual r_i by regressing X_i on all other candidate parents, $\text{PA}_i \setminus \{X_j\}$. Then, this test for conditional independence is performed by testing for independence between the residual r_i and the full set of candidate parents PA_i. If the null hypothesis of independence cannot be rejected, i.e., that the p-value is greater than a significance threshold α, the edge $X_j \to X_i$ is pruned. This procedure is applied iteratively to recover the final causal DAG.

4. EXPERIMENTS

To empirically validate the performance of our proposed beam search-based RESIT algorithm, we designed a series of simulation experiments. The primary objective is to assess the robustness and accuracy of our method compared to the standard greedy search, particularly in settings with and without unmeasured confounders. The presence of such confounders constitutes a violation of the causal sufficiency assumption, providing a key test for our algorithm's robustness.

4.1. Data Generation

We simulated data from non-linear ANMs using a data generation process that follows these steps:

1. **Graph Structure:** A random DAG is first generated over a set of nodes consisting of d observed variables and a specified number of unmeasured confounder variables.

Algorithm 1 RESIT with Beam Search

Input: Data $\mathbf{X}$, beam width w, significance α
Output: Causal order $\mathbf{O}$, Adjacency matrix $\mathbf{B}$
procedure BEAMSEARCH-RESIT($\mathbf{X}, w, \alpha$)
  $\mathbf{O}_{\text{rev}} \leftarrow$ FINDREVERSEORDER($\mathbf{X}, w$)
  $\mathbf{O} \leftarrow$ REVERSE($\mathbf{O}_{\text{rev}}$)
  $\mathbf{B} \leftarrow$ PRUNEEDGES($\mathbf{X}, \mathbf{O}, \alpha$)
  return $\mathbf{O}, \mathbf{B}$
end procedure

procedure FINDREVERSEORDER($\mathbf{X}, w$)
  $S_0 \leftarrow \{X_1, \ldots, X_d\}$
  $\mathcal{B} \leftarrow \{(\text{path} = [S_0], \text{cost} = 0)\}$ ▷ Initialize beam
  for $k = 1$ to d **do**
    $\mathcal{C} \leftarrow \emptyset$ ▷ Initialize candidates
    for all $(\pi, c) \in \mathcal{B}$ **do**
      $S_k \leftarrow$ LAST(π)
      for all $X_j \in S_k$ **do**
        $\tilde{S}_k \leftarrow S_k \setminus \{X_j\}$
        $r_j \leftarrow X_j - \hat{f}(\tilde{S}_k)$
        $\Delta c \leftarrow \text{HSIC}(\{r_j\}, \tilde{S}_k)$
        Add $(\pi \oplus [\tilde{S}_k], c + \Delta c)$ to $\mathcal{C}$
      end for
    end for
    $\mathcal{B} \leftarrow$ top-w paths from $\mathcal{C}$ by cost
  end for
  $\pi^* \leftarrow$ best path in final beam $\mathcal{B}$
  return order derived from π^*
end procedure

procedure PRUNEEDGES($\mathbf{X}, \mathbf{O}, \alpha$)
  $\mathbf{B} \leftarrow$ fully connected DAG from order $\mathbf{O}$
  for $k = 2$ to d **do**
    $X_k \leftarrow k$-th variable in $\mathbf{O}$
    $\text{PA}_k \leftarrow$ predecessors of X_k in $\mathbf{O}$
    for all $X_j \in \text{PA}_k$ **do**
      $\text{PA}_{\text{test}} \leftarrow \text{PA}_k \setminus \{X_j\}$
      if $\text{PA}_{\text{test}} \neq \emptyset$ **then**
        $r_k \leftarrow X_k - \hat{f}(\mathbf{X}_{\text{PA}_{\text{test}}})$
        $p_{\text{val}} \leftarrow \text{p-value}(\text{HSIC}(r_k, \mathbf{X}_{\text{PA}_k}))$
        if $p_{\text{val}} > \alpha$ **then**
          $B_{jk} \leftarrow 0$ ▷ Prune $X_j \to X_k$
        end if
      end if
    end for
  end for
  return $\mathbf{B}$
end procedure

We used the Erdős-Rényi model [12] to create the initial graph over the total set of nodes, with an average of 1.5 edges per node. Acyclicity is enforced by preserving only the edges that respect a random topological ordering.

2. **Causal Mechanisms:** The functional relationship for each variable is defined by a Gaussian Process (GP). This provides a flexible and powerful way to generate diverse non-linear functions. Specifically, we used a Matérn kernel, whose smoothness parameter ν controls the complexity of the functions. For our experiments, we set $\nu = 2.5$ to generate relatively smooth functions. For root nodes (those without parents), the mechanism was simplified to drawing samples from a standard normal distribution.
3. **Data Sampling:** Following the topological order of the full DAG (including the unmeasured confounders), we sample data for each variable according to its specified causal mechanism. The data corresponding to the unmeasured confounder nodes is then discarded to create the final observational dataset. This process ensures that any confounding relationships in the true graph are reflected in the statistical dependencies of the observed data.

4.2. Experimental Parameters

Table 1. Experimental Parameter Grid

Parameter	Values
Number of Nodes (d)	$\{10, 30\}$
Number of Samples (n)*	$\{100, 250, 500, 1000\}$
Confounding Level	$\{0\%, 10\%\}$ of d
Beam Width (w)*	$\{1, 2, 4, 8, 16, 32\}$

*All beam widths (w) were tested at $n = 250$ for the main analysis. Sample sizes $n = \{100, 500, 1000\}$ were evaluated only at $w = 1$ (greedy) and $w = 32$ (beam) for the robustness analysis.

Our experimental evaluation systematically assesses our method's performance in two phases. First, to analyze the primary effect of beam width (w), we ran experiments varying w from 1 to 32 for systems with $d = 10$ and $d = 30$ nodes, each under two confounding levels. This main analysis was conducted at a fixed sample size of $n = 250$. Second, to analyze the algorithm's robustness to statistical noise from finite samples, we compared the greedy search ($w = 1$) against our beam search ($w = 32$) across a wide range of sample sizes ($n = 100, 250, 500, 1000$). The complete grid of all parameters explored is summarized in Table 1. To ensure statistical robustness, we performed 50 independent trials for each unique parameter combination, using a different random seed for each trial.

4.3. Model Configuration

For all residual calculations, we employed a Gradient Boosting Regressor with 200 estimators, a learning rate of 0.2, and a maximum depth of 8. We used the Hilbert-Schmidt Independence Criterion (HSIC) as our measure of statistical dependence, both for calculating the step costs in the beam search and for the conditional independence tests in the edge pruning phase. For this final pruning phase (Algorithm 1, PRUNEEDGES), the significance level for the HSIC test was set to $\alpha = 0.01$, and an edge was removed if the p-value of its corresponding conditional independence test exceeded this threshold.

4.4. Evaluation Metrics

Our evaluation focuses on the accuracy of the final graph structure, which is the primary goal of causal discovery. We therefore use metrics that evaluate the structural accuracy of the DAG rather than the causal order itself. For any graph that is not fully connected, multiple topological orderings can be valid, so penalizing an algorithm for finding a correct but different ordering is inappropriate. Instead, we employ two standard and complementary metrics for structural accuracy: the **Structural Hamming Distance (SHD)** and the **Structural Intervention Distance (SID)**.

The SHD provides a purely structural assessment, counting the number of edge additions, deletions, or reversals required to transform the estimated graph $\hat{\mathcal{G}}$ into the true graph $\mathcal{G}$. While simple, the SHD treats all edge errors equally.

To provide a more causally-meaningful evaluation, we also use the SID [13]. This metric evaluates an estimated graph based on its ability to correctly predict the outcomes of interventions. To calculate the causal effect of an intervention on a variable i on an outcome j, one must adjust for confounding variables by controlling for a set of co-variates known as an adjustment set. A set is considered a *valid adjustment set* if it blocks all non-causal "backdoor" paths between i and j. The SID counts the number of pairs (i, j) for which the set of parents of i in the estimated graph, $\text{PA}_i^{\hat{\mathcal{G}}}$, is *not* a valid adjustment set in the true graph $\mathcal{G}$. A lower SID therefore indicates that the estimated graph is more reliable for making accurate causal predictions.

5. RESULTS

Our experimental results are organized as follows. First, in Section 5.1, we analyze the primary contribution: the trade-off between structural accuracy and computational cost as a function of the beam width (w), based on the $n = 250$ sample-size condition. Second, in Section 5.2, we analyze the algorithm's robustness to statistical noise by comparing the greedy ($w = 1$) and beam search ($w = 32$) methods across all sample sizes ($n = 100$ to 1000).

5.1. Performance and Cost vs. Beam Width

The results of our simulation experiments, summarized in Fig. 2, Fig. 3, and Table 2, confirm that the proposed beam search effectively addresses the limitations of a standard greedy search ($w = 1$). As hypothesized, we observe a consistent trend where expanding the search space via a wider beam leads to improved structural accuracy across all tested conditions.

Fig. 2 illustrates this primary finding. As the beam width (w) increases, both the SHD and SID consistently decrease, indicating that by maintaining multiple candidate orderings, the algorithm is better able to navigate the search space and avoid the irreversible errors characteristic of a greedy approach.

The advantages of a wider beam are most critical in systems with more nodes, as detailed in Table 2. For graphs with $d = 30$ nodes, the beam search ($w = 32$) achieves a highly significant improvement in SHD over the greedy

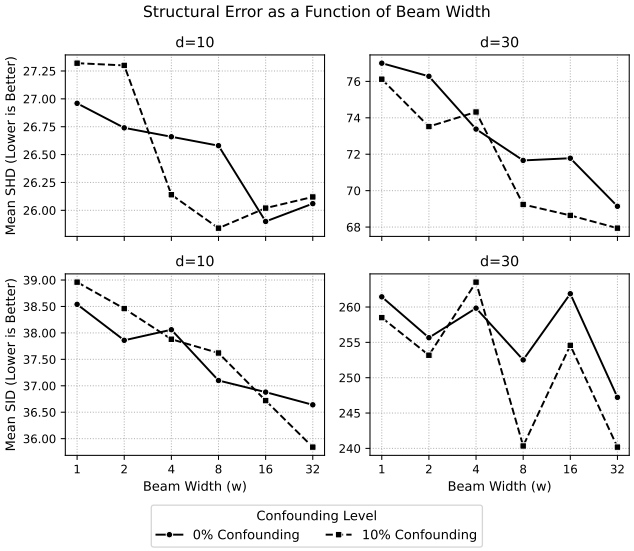

Fig. 2. Mean structural error as a function of beam width (w). The top row shows the SHD and the bottom row shows the SID for graphs with $d = 10$ and $d = 30$ nodes, under 0% confounding (solid lines) and 10% confounding (dashed lines). Each data point represents the mean over 50 independent trials. Lower values indicate better performance.

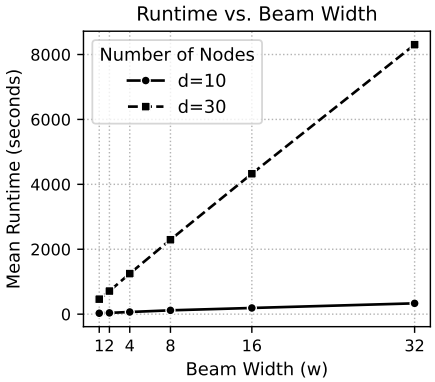

Fig. 3. Mean computational runtime as a function of beam width (w) for systems with $d = 10$ and $d = 30$ nodes. Each data point represents the average runtime over 50 independent trials.

search ($w = 1$). Specifically, SHD was reduced by **10.2%** ($p = 0.0008$) in the unconfounded case (from 77.0 to 69.1) and by **10.7%** ($p = 0.0012$) in the confounded case (from 76.1 to 67.9). The improvements in SID, which more directly measures the utility of the learned graph for causal inference, are also statistically significant. We observe an **8.0%** ($p = 0.0130$) improvement for 10-node confounded case (from 39.0 to 35.8) and a **7.1%** ($p = 0.0366$) improvement for 30-node confounded graphs (from 258.5 to 240.2). These results confirm that the performance gains from beam search are statistically robust, offering a reliable advantage especially in the challenging setting of having more nodes.

Of course, this increased accuracy is not without cost. As shown in Fig. 3, the mean runtime increases with both the

Table 2. Comparison of Greedy Search ($w = 1$) vs. Beam Search ($w = 32$). Performance is reported as Mean $\pm$ Standard Deviation over 50 trials. Significance of the improvement is denoted by asterisks on the Beam Search result: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

		SHD (↓)			SID (↓)		
Nodes (d)	Confounding	Greedy	Beam Search	Improv.	Greedy	Beam Search	Improv.
10	0%	27.0 ± 5.2	26.1 ± 5.0	3.3%	38.5 ± 6.9	36.6 ± 7.3	4.9%
	10%	27.3 ± 5.7	26.1 ± 6.0	4.4%	39.0 ± 5.7	$35.8 \pm 7.9^{*}$	8.0%
30	0%	77.0 ± 12.4	$69.1 \pm 11.7^{***}$	10.2%	261.4 ± 51.8	247.2 ± 45.1	5.4%
	10%	76.1 ± 13.4	$67.9 \pm 12.9^{**}$	10.7%	258.5 ± 52.3	$240.2 \pm 48.8^{*}$	7.1%

number of nodes (d) and the beam width (w). Crucially, the runtime scales approximately linearly with the beam width w, as the algorithm must evaluate and store up to w paths at each step. This illustrates the fundamental trade-off between accuracy and computational feasibility: the greedy search ($w = 1$) is fastest but least accurate, while a wider beam provides greater accuracy for a manageable, linear increase in computational cost. Our proposed method makes this trade-off explicit and tunable, allowing a researcher to balance the need for a robust search against practical constraints on time and resources.

5.2. Robustness to Sample Size

Table 3. Robustness to Finite Sample Sizes. SID improvement (%) of Beam ($w = 32$) over Greedy ($w = 1$). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

		SID Improv. (%) by n			
d	Conf.	100	250	500	1000
10	0%	1.8%	4.9%	$5.2\%^{*}$	−0.5%
	10%	3.0%	$8.0\%^{*}$	0.0%	1.1%
30	0%	−1.4%	5.4%	$4.7\%^{**}$	0.7%
	10%	1.1%	$7.1\%^{*}$	$6.7\%^{***}$	$1.4\%^{**}$

To analyze the algorithm's robustness to statistical noise from finite samples, we compared the greedy ($w = 1$) and beam ($w = 32$) searches across all sample sizes. We focus this analysis on the SID, as it is the more causally-meaningful metric (as discussed in Section 4.4) and best reflects the practical utility of the learned graph. The results are presented in Table 3, which also includes the $n = 250$ data from Table 2 for a comprehensive comparison.

The results reveal a nuanced relationship between sample size and the relative advantage of beam search. The most substantial and statistically significant gains are observed in the intermediate regimes of $n = 250$ and $n = 500$. At $n = 250$, beam search achieves a significant **8.0%** ($p = 0.0130$) SID improvement in the 10-node confounded case and **7.1%** ($p = 0.0366$) in the 30-node confounded case. This advantage is even more pronounced at $n = 500$, yielding highly significant improvements of **5.2%** ($p = 0.0151$), **4.7%** ($p = 0.0063$), and **6.7%** ($p < 0.0001$) across various conditions. This suggests an optimal range for our method: at these intermediate sample sizes, the data contains enough signal to find the true causal structure, but also contains sufficient statistical noise to trap a purely greedy ($w = 1$) algorithm in a local optimum. The wider search of our beam ($w = 32$) approach is able to successfully navigate this noise and identify a better causal ordering.

Conversely, this relative advantage diminishes at the extremes. At the smallest sample size ($n = 100$), the data is likely dominated by statistical noise. The improvements are modest and not statistically significant (e.g., 3.0% where $p = 0.2120$ and 1.1% where $p = 0.3512$ in confounded cases), suggesting that neither method can reliably find the true structure. As the sample size becomes large ($n = 1000$), the performance of the greedy search largely converges with that of the beam search. With plentiful, reliable data, the local optimum found by the greedy search is often the global optimum, and the additional computational effort provides diminishing returns (e.g., 1.1% or −0.5% improvement). We do note one exception in the 30-node confounded case, where a small but highly significant improvement of 1.4% ($p = 0.0026$) remains, suggesting that even with large data, beam search can still offer a modest advantage in highly complex scenarios. Overall, this demonstrates that the beam search is a particularly valuable tool in intermediate, high-variance regimes where a greedy search is most likely to fail, but where enough signal exists to make a more robust search worthwhile.

6. DISCUSSION

In this paper, we addressed a fundamental limitation of greedy search algorithms in causal discovery: their susceptibility to making irreversible errors, especially when local errors occur, whether from violations of the core assumption of causal sufficiency or from statistical noise from finite samples. Our goal was to develop a more robust method for learning ANMs by generalizing the local search of RESIT into a more comprehensive, yet still practical, search strategy. The proposed beam search algorithm was designed to achieve this by exploring multiple promising causal orderings simultaneously, thereby mitigating the risk of committing to an early, erroneous decision.

Our empirical results provide strong validation for this approach. As hypothesized, widening the search beam consistently led to more accurate causal graphs, with the most significant gains observed in the very scenarios where greedy methods are most likely to fail: systems with more nodes and those with unmeasured confounders. Furthermore, our analysis across all sample sizes (Section 5.2) reveals that the gains from beam search are most significant in intermediate, high-variance regimes ($n = 250, 500$), as shown in Table 3.

This suggests an optimal sample size range where the data is noisy enough to trap a greedy search but contains enough signal for our robust method to exploit. The statistically significant improvements in both SHD and SID demonstrate that our method is not only structurally more accurate but also produces models that are more reliable for downstream causal inference tasks. This directly addresses the core problem outlined in our introduction. By framing the task as a path search problem and using beam search as a tractable approximation, we have successfully enhanced the robustness of ANM-based discovery without resorting to a computationally infeasible exhaustive search.

This work also highlights the critical trade-off between search completeness and computational cost. The beam width, w, serves as an explicit and tunable parameter that allows researchers to navigate this trade-off. While a greedy search ($w = 1$) offers speed, our findings show that even a modest increase in beam width can yield substantial accuracy improvements. Crucially, this increase in runtime scales linearly with w, representing a manageable, practical cost for the enhanced robustness. This provides a practical path forward for tackling more complex, real-world problems where greedy search may fail.

Despite these promising results, our method has several limitations that open avenues for future research. First, while more robust than a greedy search, beam search is still a heuristic. It is not guaranteed to find the globally optimal causal ordering and can still be trapped in local optima, albeit on a larger scale. The performance of the algorithm also fundamentally depends on the chosen regression model and independence test. Our use of Gradient Boosting and HSIC proved effective, but exploring more advanced or specialized components could yield further improvements.

Furthermore, this study was conducted on simulated data. While this provides a controlled environment for rigorous evaluation, the ultimate test of any causal discovery method is its application to real-world observational data. Future work should focus on applying this beam search framework to complex datasets from fields such as robotics, economics, or systems biology, where unmeasured confounding is a pervasive and critical issue. This would likely require integrating more sophisticated model selection techniques to handle the noise and heterogeneity inherent in such data. Another promising direction is the exploration of adaptive beam strategies, where the algorithm could dynamically adjust the beam width based on the uncertainty or complexity of the current step in the search.

7. CONCLUSION

In conclusion, by generalizing the greedy search of RESIT to a beam search, we have developed a powerful and practical framework for causal discovery in ANMs. Our method offers a significant and tunable improvement in robustness over traditional approaches, making it a valuable tool for more reliable causal discovery from observational data, especially in the challenging, non-ideal conditions often encountered in scientific and engineering.

ACKNOWLEDGMENTS

This study is supported by the Japan Society for the Promotion of Science through the Grants-in-Aid for Scientific Research Program (KAKENHI 18K19821 and 22K19833) and by Kyoto University and Toyota Motor Corporation through the joint project Advanced Mathematical Science for Mobility Society.

REFERENCES

[1] B. Scholkopf, F. Locatello, S. Bauer, N.R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward Causal Representation Learning," Proc. IEEE, vol.109, no.5, pp.612–634, May 2021.

[2] J. Pearl, Causality: models, reasoning, and inference, Cambridge University Press, Cambridge, U.K. ; New York, 2000.

[3] J. Peters, D. Janzing, and B. Schölkopf, Elements of causal inference: foundations and learning algorithms, Adaptive computation and machine learning series, The MIT Press, Cambridge, Massachuestts, 2017.

[4] J. Peters, J.M. Mooij, D. Janzing, and B. Schölkopf, "Causal discovery with continuous additive noise models," Journal of Machine Learning Research, vol.15, no.58, pp.2009–2053, 2014.

[5] S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-gaussian acyclic model for causal discovery," J. Mach. Learn. Res., vol.7, p.2003–2030, Dec. 2006.

[6] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P.O. Hoyer, and K. Bollen, "Directlingam: A direct method for learning a linear non-gaussian structural equation model," J. Mach. Learn. Res., vol.12, no.null, p.1225–1248, July 2011.

[7] J. Suzuki and T.L. Yang, "Generalization of lingam that allows confounding," 2024 IEEE International Symposium on Information Theory (ISIT), pp.3540–3545, 2024.

[8] Y. Sogawa, S. Shimizu, Y. Kawahara, and T. Washio, "An experimental comparison of linear non-gaussian causal discovery methods and their variants," The 2010 International Joint Conference on Neural Networks (IJCNN), pp.1–8, 2010.

[9] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Scholkopf, and A. Smola, "A kernel statistical test of independence," Neural Information Processing Systems, 2007.

[10] T. Ikeuchi, M. Ide, Y. Zeng, T.N. Maeda, and S. Shimizu, "Python package for causal discovery based on lingam," Journal of Machine Learning Research, vol.24, no.14, pp.1–8, 2023.

[11] S.J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach (4th Edition), Pearson, 2020.

[12] E.N. Gilbert, "Random graphs," Annals of Mathematical Statistics, vol.30, no.4, pp.1141–1144, 1959.

[13] J. Peters and P. Bühlmann, "Structural intervention distance for evaluating causal graphs," Neural Comput., vol.27, no.3, p.771–799, March 2015.