G-RIPS Sendai 2026

Fujitsu Group

Final Report

Authors:
Brandin Farris[1]
Misaki Nagashima[2]
Wu Tewei[3]
Jonah Yoshida[4]

Mentors:
Masaki Ogawa[3]
Eriko Shinkawa[3]

[1] Oregon State University
[2] Hokkaido University
[3] Tohoku University
[4] University of California, LA

August 6, 2026

Abstract

Fujitsu Causal AI is an algorithm/pipeline that takes a large dataset as input and outputs a collection of conditional causal graphs with the intent of developing a causal discovery pipeline from a single dataset. In this project, we develop novel methods for analyzing the output of Fujitsu Causal AI that are motivated by geometry and topology, and apply those methods to materials science simulation data and synthetic data to showcase its potential. In particular, we develop a new algorithm that clusters the output of Fujitsu Causal AI in a way that respects the topology of the underlying spaces. We also develop theory involving direct LiNGAM as a function from condition space to causal graph space.

Contents

1 Introduction **3**
1.1 Background . . . 3
1.2 Fujitsu Causal AI . . . 3
1.3 Problem Statement . . . 3
1.4 Core Requirements for Practical Causal Discovery . . . 3

2 Choosing our Metric **4**
2.1 Metrics on Causal Graphs . . . 4
2.2 Preliminary Results on TEM . . . 6
2.3 Metrics on Conditions . . . 7

3 The Pathing Algorithm **8**
3.1 Condition-Based Pathing and Interpretations . . . 9

4 Theoretical Guarantees **10**
4.1 Stability . . . 10

5 Topology of Causal Graph Space **16**
5.1 Why Study Topology? . . . 16
5.2 The Topology of the Causal Graph Space . . . 16

6 Experiments **18**
6.1 Choosing a dataset . . . 18
6.2 Experiments for Conditional Pathing . . . 19
6.3 Experiments for Winding Around Cycle Lines . . . 22

7 Conclusion **22**

8 Future Work **24**
8.1 Causal Graph Topology Conjectures . . . 24
8.2 Optimal Pathing . . . 26
8.3 Picking (ϵ, δ) for pathing, filtrations . . . 26
8.4 The conjectures of modulus continuity . . . 27
8.5 Formalization of the condition space and data space . . . 27

A Caveats to Conditional Direct LiNGAM **29**

B Calculation details in theoretical guarantee **29**

C Contributions **33**
C.1 Misaki Nagashima . . . 33
C.2 (Brandin) Farris . . . 34
C.3 Wu TEWEI . . . 34
C.4 Jonah Yoshida . . . 34

References **34**

It is not that we have a short time to live, but that we waste much of it.

Seneca

1 Introduction

1.1 Background

In recent years, advances in both experimental and simulation techniques have significantly improved the resolution and volume of data in materials science.

On the experimental side, for example, the launch of NanoTerasu [12]—a next-generation synchrotron light source facility in Japan (2024)—has enabled high-precision observation of molecular, atomic, and electronic states at the nanometer scale. Concurrently, in computational science, AI-assisted molecular dynamics (MD) simulations allow for the highly accurate reproduction of chemical reactions and atomic behaviors in large-scale systems exceeding 100,000 atoms.

With high-resolution data increasingly generated from both physical experiments and theoretical simulations, materials science has entered a data-rich era that creates new opportunities for data-driven analysis. Experimental and computational advances have generated high-dimensional datasets that exceed human processing capacity. Relying solely on intuition or prior knowledge to analyze such datasets risks missing unexpected insights. Thus, the core challenge: extract meaningful knowledge from complex datasets while minimizing human bias. Causal AI addresses this challenge by aiming to infer underlying causal mechanisms rather than mere correlations.

1.2 Fujitsu Causal AI

A core component of Causal AI is Causal Discovery, which automatically extracts causal relationships from observational data. Conventional algorithms, such as DirectLiNGAM (Direct Linear Non-Gaussian Acyclic Model), estimate a single Directed Acyclic Graph (DAG) to represent the overall causal structure. However, in real-world materials data, causal mechanisms are rarely static and can vary significantly depending on physical or chemical conditions (e.g., variations in temperature, pressure, or atomic concentration). Consequently, relying on a single overall graph risks oversimplifying condition-dependent phenomena.

To address this limitation, this study adopts Fujitsu Causal AI—a framework that integrates Wide-Learning™ (a rule-based knowledge discovery technique made by Fujitsu) with DirectLiNGAM[14] —as our baseline approach. In this framework, Wide-Learning™ exhaustively extracts sub-conditions (condition rules) from the dataset, and DirectLiNGAM is subsequently applied to the data subset under each condition. This enables the exhaustive estimation of distinct causal graphs tailored to different conditions.

1.3 Problem Statement

However, this exhaustiveness introduces a practical challenge in managing the resulting multi-graph ensemble. Because subtle variations in conditions can generate thousands of distinct causal graphs, interpreting such a large collection of models becomes highly complex. Relying on domain intuition to manually select or aggregate these graphs risks introducing subjective human bias into the evaluation process.

Systematically interpreting this multi-graph ensemble requires analyzing the structural relationships among condition-specific sub-graphs without subjective filtering. In this work, we address this issue by introducing novel metrics on the causal graph space to evaluate sub-graph connectivity and track how causal structures transition across changing conditions. Furthermore, by integrating this pathing analysis with a topological characterization of the graph space, we aim to construct a unified "Global Causal Landscape."

1.4 Core Requirements for Practical Causal Discovery

To connect complex graph ensembles with practical decision-making, our framework is designed around three key principles for real-world application:

1. **Convincingness:** Transform complex graph ensembles into simple and interpretable structures that domain experts can readily evaluate and accept.

2. **Comprehensiveness:** Preserve the exhaustive analysis capabilities of Wide-Learning™ to extract insights from data without being limited by prior human assumptions.

3. **Discoverability:** Identify novel or non-obvious patterns that may initially lack intuitive convincingness, yet represent valuable scientific or operational insights.

2 Choosing our Metric

2.1 Metrics on Causal Graphs

In order to view the data space as a metric space, which has nice Hausdorff and completeness properties, we want to define a metric between causal graphs. Additionally, having a metric space allows us to naively perform TDA on any data. We look first at a few candidates and detail why our final choice is both a metric on causal graphs (directed acyclic graphs, DAGs) and captures desirable causal information.

- **SHD**: The Structural Hamming Distance (SHD) is defined as

  $$\text{SHD}(G, H) = \min(\text{number of edge additions, deletions, and reversals transforming G into H}).$$

  For instance, $\text{SHD}(G, H) = 2$ for $G : X \to Y, Y \to Z$ and $H : Z \to Y$, with one deletion and one arrow-reversal.

- **SID**: The Structural Intervention Distance Metric (SID) is defined as

  $$\text{SID}(G, H) = |\{(X, Y) \in V^2 : X \neq Y, \text{Pa}_H(X) \notin \mathcal{A}_G(X, Y)\}|,$$

  where $\mathcal{A}_G(X, Y)$ denotes the collection of valid adjustment sets for estimating the causal effect of X on Y in G. We give a rigorous definition of $\mathcal{A}_G(X, Y)$ by defining colliders.

  Definition 1. A *collider* V_k on a path from X to Y is such that $V_i \to V_k \leftarrow V_j$ for some V_i, V_k, V_j in order on the path. A non-collider is a point on the path that is not a collider. For some set Z, a non-collider V_k *blocks* the path if $V_k \in Z$. A collider V_k blocks the path if $V_k \notin Z$ and no descendant of V_k is in Z.

  Definition 2.

  $$Z \in \mathcal{A}_G(X, Y) \iff \begin{cases} Z \text{ does not block any causal } X \to Y \text{ path,} \\ Z \text{ blocks every noncausal path from } X \text{ to } Y. \end{cases}$$

Example 1. We now compare SID and SHD to demonstrate an instance in which SID is preferred. Let $G : Z \to X \to Y, Z \to Y$ and $H : X \to Z \to Y, X \to Y$. By one arrow-reversal, $\text{SHD}(G, H) = 1$. Now, the parent sets in H are $\text{Pa}_H(X) = \{Z\}, \text{Pa}_H(Z) = \{X\}, \text{Pa}_H(Y) = \{X, Z\}$. Computing $\mathcal{A}_G(*, *)$:

$$\mathcal{A}_G(X, Y) = \{\{Z\}\}, \qquad \mathcal{A}_G(Y, X) = \{\{Z\}\}, \qquad \mathcal{A}_G(X, Z) = \varnothing,$$

$$\mathcal{A}_G(Z, X) = \{\varnothing\}, \qquad \mathcal{A}_G(Z, Y) = \{\varnothing\}, \qquad \mathcal{A}_G(Y, Z) = \varnothing.$$

So, the only contributions to the SID sum come from $X \in \text{Pa}_H(Z), X \in \text{De}_G(Z)$ on the pair (Z, X) and $Y \notin \text{Pa}_H(Z)$ with $X \notin \mathcal{A}_G(Z, Y)$ as it blocks $Z \to X \to Y$, on the pair (Z, Y). Thus, $\text{SID}(G, H) = 2$, giving a larger distance than SHD and capturing a larger change in downstream effects from the change in top causal order of nodes between G and H, from (Z, X, Y) to (X, Z, Y).

Although SID captures these causal chain changes well, it is only a metric on adjacency matrices, not weighted adjacency matrices. As we are interested in the magnitude of weights that measure the strength of causal relationships, one might try the contSID metric between DAGs G, H formulated by Dhanakshrirur, Laumann, Park, and Barahona in 2024 [5]. Still, this metric requires an underlying observational distribution P and is not a pure metric on the set of DAGs alone. Thus, we have motivation for a total effect metric which 1) captures causal effect chains and 2) does not require assumptions about observational distributions.

- **TEM**: The Total Effect Metric (TEM) between two acyclic graphs G, H on the same node set is defined by first taking their weighted adjacency matrices G', H'. Then,

$$d_{TE}(G,H) = ||\mathcal{F}(G') - \mathcal{F}(H')||_F,$$

where $||\cdot||_F$ is the Frobenius metric and $\mathcal{F} : M \mapsto (I-M)^{-1} - I = \sum_{i=1}^{\infty} M^i$ is well-defined as G' is nilpotent, so the sum terminates. Since $M^i_{j,k}$ counts the sum of products of weights for i-length paths from j to k, $(\sum_{i=1}^{\infty} M^i)_{j,k}$ counts the sum of products of weights for all paths from j to k.

Example 2. Consider the weighted causal graph $G : X \xrightarrow{2} Y \xrightarrow{1} Z, X \xrightarrow{1} Z$. Consider the same graph with the top arrow reversed: $H : X \xleftarrow{2} Y \xrightarrow{1} Z, X \xrightarrow{1} Z$, represented by matrices $G' = \begin{pmatrix} 0 & 2 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, H' = \begin{pmatrix} 0 & 0 & 1 \\ 2 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ on the canonical node ordering (X, Y, Z). Then, $\sum_{i=1}^{\infty} G'^i = \begin{pmatrix} 0 & 2 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \sum_{i=1}^{\infty} H'^i = \begin{pmatrix} 0 & 0 & 1 \\ 2 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix}$ with $d_{TE}(G,H) = ||\mathcal{F}(G') - \mathcal{F}(H')||_2$ $= \sqrt{(3-1)^2 + (2-0)^2 + (0-2)^2 + (1-3)^2} = 4$. Now, consider the graph G with the bottom arrow reversed and do the same distance-to-G calculation: $N : X \xrightarrow{2} Y \xleftarrow{1} Z, X \xrightarrow{1} Z$ with $\sum_{i=1}^{\infty} N'^i = \begin{pmatrix} 0 & 3 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, then $d_{TE}(G,N) = ||F(G') - F(N')||_F = \sqrt{1+1+1+2^2} = \sqrt{7} < 4$. Indeed, reversing arrows corresponding to smaller lexicographic changes in the causal order gives smaller changes in the total effect metric.

In hopes of better interpreting the distances between causal graph pairs, we choose TEM in this paper. It is of interest to consider other related metrics, especially those that respect causal chain structures, such as contractions on Markov Equivalence Classes (MECs).

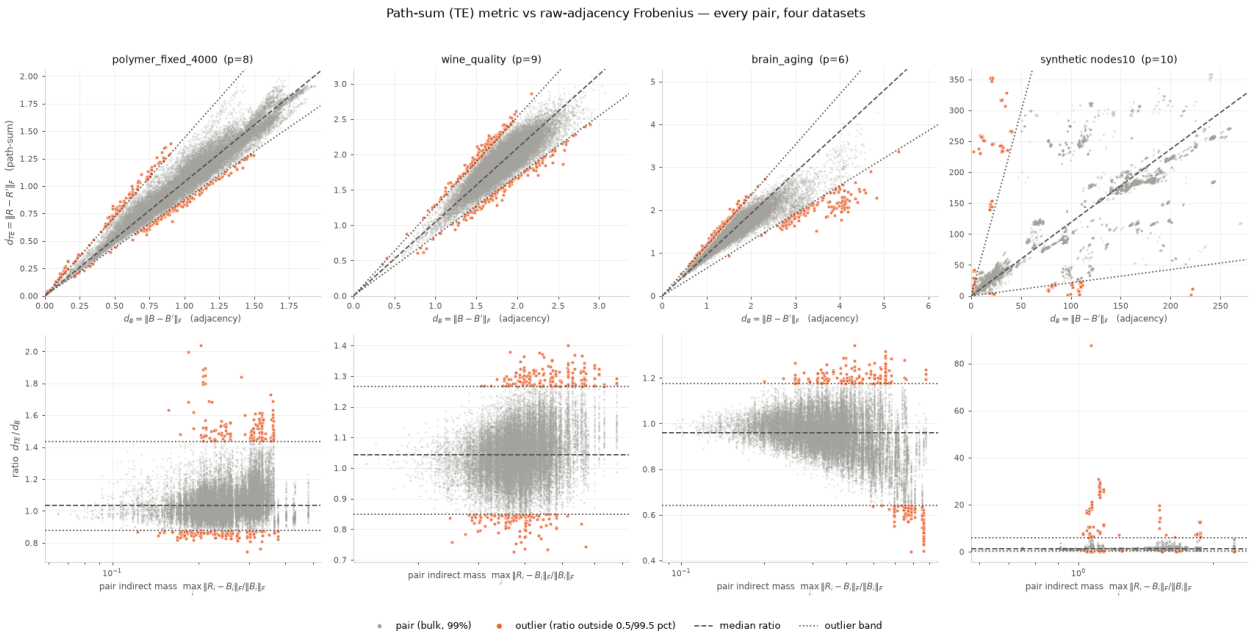


Figure 1: Plots of TEM with (y-axis) and without (x-axis) the F map on data. Orange points are outliers.

2.2 Preliminary Results on TEM

Lemma 1. *$\mathcal{F} : CGS \to \mathbb{R}^{n\times n}$ is an injective function.*

Proof. Let $\mathcal{F}' : CGS \to M$ where $\mathcal{F}'$ takes DAGs to their strictly upper triangular weighted adjacency matrices per a fixed topological ordering. As $\mathcal{F}'$ is bijective, it suffices to show that $f : M \to \mathbb{R}^{n\times n}$ is injective, where $f \circ \mathcal{F}' = \mathcal{F}$. It turns out that $f(M) + I = \sum_{k=0}^{\infty} M^k$, and M nilpotent with $M^s = 0$ gives $f(M) + I = \sum_{k=0}^{s} M^k$ such that $(f(M) + I)(I - M) = I - M^{s+1} = I \implies f(M) = (I - M)^{-1} - I$. Thus, $f(M)$ is injective, and so is $\mathcal{F}$. □

Corollary 2. *TEM is a metric on DAGs. As $\mathcal{F}$ is injective and the ℓ_2-norm induces a metric on $M_n(\mathbb{R})$, we have an induced metric on CGS, namely TEM.*

Lemma 3. *$\mathcal{F} : CGS \to SRT \subset \mathbb{R}^{n\times n}$ is a surjective function, where SRT are strictly upper triangular matrices.*

Proof. Let our SRT matrix be S. Consider the full DAG F_n consisting of vertices $1, 2, \ldots, n$ with directed edge $(i, j) \iff i < j$. Then, $\mathcal{F}(F_n)_{(i,j)} = \sum_{p\in P_i(j)} c(p) = \sum_{k=i+1}^{j-1} \sum_{\ell_1\in P_i(k)} \sum_{\ell_2\in P_k(j)} c(\ell_1)c(\ell_2) + w_{(i,j)}$. As $\sum_{k=i+1}^{j-1} \sum_{\ell_1\in P_i(k)} \sum_{\ell_2\in P_k(j)} c(\ell_1)c(\ell_2)$ is some homogeneous equation in $w_{(q,w)}$ where $w - q < j - i$, we can let $w_{(i,j)} = s_{(i,j)}$ for $j = i + 1$, $w_{(i,j)} = s_{(i,j)} - h_{(i,j)}(w_{(k,k+1)\forall i\le k<j})$ for $j = i + 2$, and so on such that $\mathcal{F}(F_n) = S$. □

Example 3. Consider the upper triangular $S = \begin{pmatrix} 0 & 2 & 5 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix}$. Then, our full DAG F_3 will have edge set $1 \to 2, 2 \to 3, 1 \to 3$ with $w_{(1,2)} = 2, w_{(2,3)} = 3, w_{(1,3)} = s_{(1,3)} - w_{(1,2)}w_{(2,3)} = -1$ so that $\mathcal{F}(F_3) = S$.

Lemma 4. *(CGS, d) with $d(G, G') = ||\mathcal{F}(G) - \mathcal{F}(G')||_F$ is a complete metric space.*

Proof. As $(M_n(\mathbb{R}), F)$ is a complete metric space, it suffices to show that $\mathrm{Im}(f)$ is a closed subset of $M_n(\mathbb{R})$, where $f : S := \{M \in M_n(\mathbb{R}) | \forall k \forall (i, j) M^k_{(i,i)} = 0 \wedge (M_{(i,j)} \neq 0 \implies M_{(j,i)} = 0)\} \cong CGS \to M_n(\mathbb{R}); M \mapsto \sum_{i=1}^{\infty} M^i$. As $(M_n(\mathbb{R}), F)$ is Hausdorff by choosing $\epsilon = \frac{1}{3} \sup_{(i,j)\in[n]^2} |G_{(i,j)} - H_{(i,j)}| \; \forall G, H \in M_n(\mathbb{R})$ so that $B_\epsilon(G_1) \cap B_\epsilon(G_2) = \emptyset$, we show that f is a proper map to conclude $\mathrm{Im}(f)$ closed. First, consider the map $f_1 : S \to U_n(\mathbb{R}); M \mapsto I - M$, to unipotent matrices. Consider $K \subset U_n(\mathbb{R})$ compact. Then, $f_1^{-1}(K) = \{M \in S | I - M \in K\} = S \cap (I - K)$. As $M \mapsto I - M$ is a homeomorphism between $\tilde{N} := \{N \in M_n(\mathbb{R}) \mid \text{N is nilpotent}\}$ and $U_n(\mathbb{R})$, $I - K \subseteq \tilde{N}$ must also be compact. Thus, if $S \subset \tilde{N}$ is closed, $f^{-1}(K)$ is compact, and f_1 is proper. As $\tilde{N} \subset M_n(\mathbb{R})$ is closed, we only need show that $S \subset M_n(\mathbb{R})$ is closed. Consider that S^c is the union of two sets S_1, S_2 where the first and second conditions are violated. As S_1 is a subset of cyclic graphs under the graph isomorphism, simply choose $\epsilon < \inf(w_{(i,j)})$ for a graph G's directed weights $w_{(i,j)}$ such that the ϵ-ball of the corresponding matrix of G is also a subset of the cyclic graphs (i.e. edges cannot be destroyed with a small enough perturbation). As S_2 corresponds to graphs with backward edges, the same choice of ϵ as above suffices to ensure that graphs corresponding to matrices in this ϵ-ball also have backward edges. Thus, S_1 and S_2 are open, S^c is open, and f_1 is proper. Now, consider the map $f_2 : U_n(\mathbb{R}) \to M_n(\mathbb{R}); M \mapsto M^{-1}$ and some compact $K \subset M_n(\mathbb{R})$. As $g : U_n(\mathbb{R}) \to U_n(\mathbb{R}); M \mapsto M^{-1}$ is a homeomorphism, $K \cap U_n(\mathbb{R})$ must be compact by $U_n(\mathbb{R}) \subset M_n(\mathbb{R})$ closed. Thus, $g^{-1}(K \cap U_n(\mathbb{R})) = f_2^{-1}(K)$ is also compact. In all, $f = f_2 \circ f_1 - I$, and the composition of proper maps being proper finishes the proof (i.e. the $-I$ shift map is a homeomorphism $f_3 : M_n(\mathbb{R}) \to M_n(\mathbb{R})$, thus a proper map). □

We now have a few nice properties of our TEM. In particular, completeness guarantees that we can study limits of causal graphs in our space, such as limits to the 0-graph or other graphs on $\partial(CGS)$ from $\mathrm{int}(CGS)$, causal graphs that have one node in each topological ordering layer with full filling of directed edges. Next, we transition to a bound on the distance for a small perturbation to the adjacency matrix and discuss results on real data.

Lemma 5. *Define the bijection* $CSG \cong_i S$. *Let* $G, G' \in CGS$ *such that* $i(G') = i(G) + P$ *for* $P_{(i,j)} = 0 \iff \neg\exists(i,j) \in E(G)$. *If* $\epsilon = ||P||_{\mathcal{F}} < \frac{1}{d(G,0)+\sqrt{n}}$ *for* 0 *the graph with* $E(0) = \emptyset$, *then*

$$\boxed{d(G', G) \leq \frac{\epsilon \cdot (d(G,0) + \sqrt{n})^2}{1 - \epsilon \cdot (d(G,0) + \sqrt{n})}.}$$

Proof. We appeal to triangle inequality in both Frobenius and defined metrics.
$d(G', G) = ||(I - i(G'))^{-1} - (I - i(G))^{-1}||_F$
$= ||(I - i(G'))^{-1}(I - i(G) - (I - i(G')))(I - i(G))^{-1}||_F$
$= ||(I - i(G'))^{-1}P(I - i(G))^{-1}||_F \leq ||(I - i(G'))^{-1}||_F||P||_F||(I - i(G))^{-1}||_F$
$\leq \epsilon \cdot ||(I - i(G'))^{-1} - I + I||_F||(I - i(G))^{-1} - I + I||_F$
$\leq \epsilon \cdot (||(I - i(G'))^{-1} - I||_F + ||I||_F)(||(I - i(G))^{-1} - I||_F + ||I||_F)$
$\leq \epsilon \cdot (d(G', 0) + \sqrt{n}) \cdot (d(G, 0) + \sqrt{n})$
$\leq \epsilon \cdot (d(G', G) + d(G, 0) + \sqrt{n}) \cdot (d(G, 0) + \sqrt{n})$.
Rearranging, $d(G', G)(1 - \epsilon \cdot (d(G, 0) + \sqrt{n})) \leq \epsilon \cdot (d(G, 0) + \sqrt{n})^2$. Dividing both sides by positive $1 - \epsilon \cdot (d(G, 0) + \sqrt{n})$ gives the desired result. □

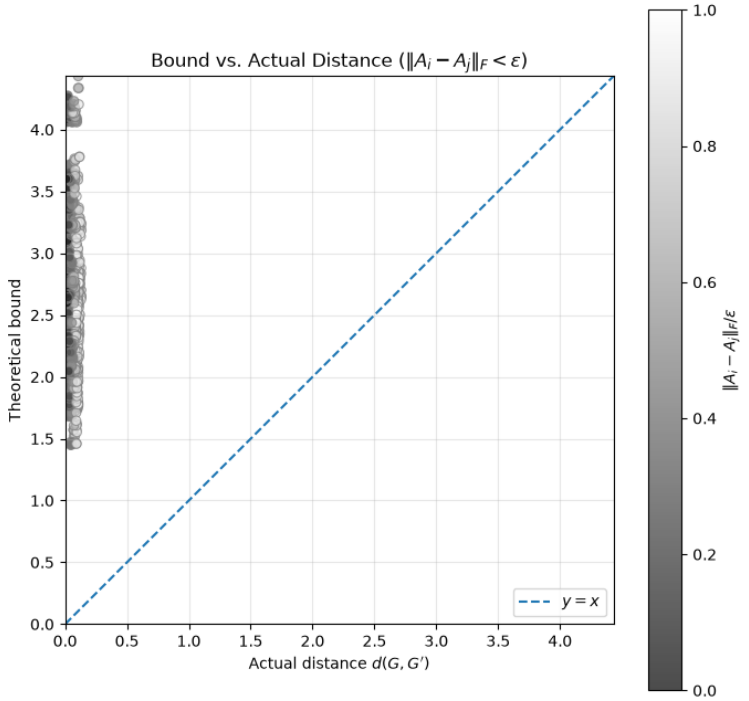


Figure 2: Bound v.s. Distance plot with $B_\epsilon(i)$ shading

2.3 Metrics on Conditions

Suppose a material scientist is experimenting with the *conductivity* of various quantum semiconductors. Then, they should expect there to be a causal graph between some of their variables that captures the ground truth of working within a *quantum regime*. Say that the variables voltage (V), current (I), temperature (T), and noise (S) are column headers. Then, Figure 4 describes the inherent physical relationships that dictate causality between these four variables in a quantum and in a classical setting.

Notice that there is a causal triangle fundamental to the quantum regime that is not present in the classical regime. Intuitively, then, the set of conditions can dictate which causal graphs DirectLiNGAM produces. It is thus of interest to look at metrics on conditions depending on the underlying data to analyze realistic

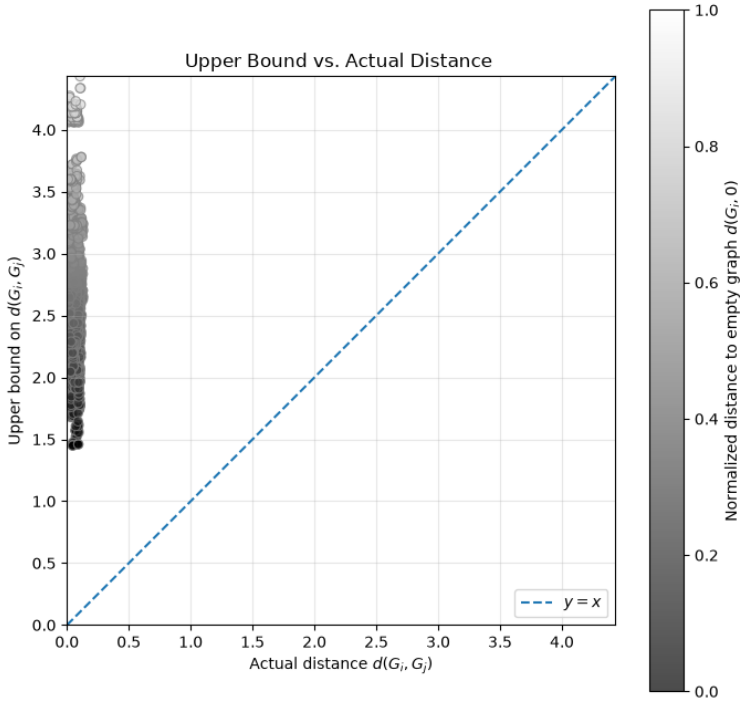
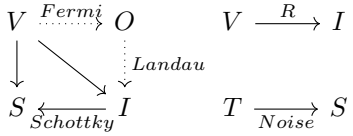


Figure 3: Bound v.s. Distance plot with $d(i, 0)$ shading

Figure 4: Sample causal graphs in a quantum (left) and a classical (right) setting. The hidden variable O represents an aggregate occupation state for the electrons that conduct charge.



causal graph structures that arise from considering data under similar conditions. Though there are several options for such metrics, we choose the Jaccard metric as it appears as the conventional condition metric in Fujitsu's causal discovery platform [8] and allows for DirectLiNGAM continuity analysis when isolated data points are added or removed. Other potential choices for a condition metric include a standard Hamming metric and a lexicographic metric.

3 The Pathing Algorithm

Given several thousand condition-graph pairs, it becomes difficult to relate and understand them in their entirety. How do we determine when two conditions lead to the "same" causal graph? Better yet, what is the correct notion of "sameness" to use when comparing condition-graph pairs? One natural choice is to apply standard clustering algorithms. Given a set of points in $\mathbb{R}^d$ and some input parameter (such as k for $k-$means clustering), partition the set in some sensible way. This is an often informative approach and simplifies the data to help researchers understand and interpret. In fact, this is what Fujitsu did in their paper introducing the causal discovery pipeline.

One problem with standard clustering is that it inherently ignores the structure of the involved spaces. The causal graph space described in 5.2 is nontrivial and complicated. In such cases, clustering purely based on pair-wise distances ignores this topology and therefore ignoring part of the relationship between condition-graph pairs. Consider a pair of points in the CGS whose line in $M_n(\mathbb{R})$ passes through a cyclic graph. Then,

the geodesic distance and Frobenius distance are different. For this reason, applying a classification method that respects the topology of the underlying space has the potential to provide more qualitative information than classical clustering techniques in the context where we have nontrivial geometry in our space, as we do here.

Another choice of sameness is path components, a notion from topology. Path components are inherently **comprehensive** because they segment the entire space (in this case, all possible DAGS that can come from a dataset), so it is an all-encompassing segmentation. The pathing algorithm is a method to compute gradually changing paths between the condition-graph pairs by changing the condition in a gradual way such that the resulting graph also changes graudally. We can do this on all pairs simultaneously, resulting in a picture capturing the path components of the output of Fujitsu's causal discovery algorithm.

There are several benefits of this approach. Path components are a deterministic topological invariant, meaning that when we treat $\mathcal{F} : \mathcal{C} \to CGS$ as a function from the underlying distribution of a dataset (i.e. assuming infinite sampling), it has a ground truth collection of path components that perfectly segment the graph of F. This is theoretical ground truth to strive for.

Additionally, we can study the paths themselves to learn information about how the condition space interacts with causal mechanisms. As an example of this, say we connect two condition-graph pairs by a path. If, along that path, there are literals (that is, specific conditions within a given condition, such as an inequality on molality) that change freely along the path, then these literals are *mechanism irrelevant candidates*, in that their contribution to the causal graph minimally changes the(or some of the) causal mechanisms along the path (example in 6.2). The literals that don't change at all throughout the path are candidates for *load bearing literals*, meaning that their condition is a primary cause of the causal mechanism in question (see 6.2).These interpretations enhance **discoverability** by showing us what conditions are important and unimportant to certain mechanisms found by Fujitsu Causal AI. These interpretations are preliminary, and assume that the path taken is reasonable, which is dependent on the choice of (ϵ, δ) and on the spread of the data in question. This is also why ϵ and δ are variables and not fixed parameters. We explore this choice further in 8.3. From our limited exploration of the paths thus far, some paths follow this pattern while others don't. Further exploration is necessary on the same dataset with resampling, and different datasets to see what patterns persist.

3.1 Condition-Based Pathing and Interpretations

We start with a definition.

Definition 3. Let (C_0, G_0) and (C_K, G_K) be two condition-graph pairs. Let $\epsilon, \delta > 0$. Then (C_0, G_0) and (C_K, G_K) are (ϵ, δ) path connected if there exists a sequence of conditions $\{C_i\}_{i=0}^{K}$ such that for each i,

1. $d_J(C_i, C_{i+1}) \leq \delta$

2. $d_{TE}(G_i, G_{i+1}) \leq \epsilon$

One may notice that this definition shares in flavor with a freshman calculus definition of continuity. While F is not continuous, our experiments have lead us to a conjecture that it is piece-wise continuous (see section 4). This definition is connecting the outputs of the causal discovery that are, in some sense, within the same area or regime of continuity. We say *some regime*, because continuity states that for every x, for every ϵ, *there exists* a δ. Without such guarantees, δ must become a variable. Note also that this definition is closer to *uniform continuity* because the choice of δ here is independent of a given condition(i.e. independent of our "x"); This is not a stretch because our conjectures are that of lipschitz-like behavior, so uniform continuity follows.

In the case where the user has two particular condition-graph pairs they wish to study, testing whether they are path connected is one informative invariant. The path(s) between two pairs can be studied as described above, giving insight into which condition literals are contributing the most to the causal mechanisms in question. If instead one wishes to consider the entire output of Fujitsu Causal AI at the same time, we can use this algorithm to find the *path components* of the output. As we will see in 6.2, path components give us a way to cluster the output in a way that respects the underlying topologies, and these clusters can be studied to understand the conditional causal mechanisms underlying the dataset.

Algorithm 1 (ϵ, δ)-Conditional Pathing (A*)

Require: conditions $C_{\text{src}}, C_{\text{tgt}}$; data X; bounds δ, ϵ; weights α, β, γ; cap $N_{\max}$
Ensure: an (ϵ, δ)-path $C_{\text{src}} \to C_{\text{tgt}}$, or NONE

$R_{\text{tgt}} \leftarrow \mathrm{F}(C_{\text{tgt}})$, $R_{\text{src}} \leftarrow \mathrm{F}(C_{\text{src}})$; **if** either $= \perp$ **return** NONE
$h(R) := \|R - R_{\text{tgt}}\|_F$; $g[\rho(C_{\text{src}})] \leftarrow 0$; $n \leftarrow 0$
push $\big(h(R_{\text{src}}),\, C_{\text{src}},\, R_{\text{src}}\big)$ onto min-heap Q ▷ keyed by $f = g + h$
while $Q \neq \emptyset$ **do**
  pop (f, C, R) least; $S \leftarrow \rho(C)$
  if $\|R - R_{\text{tgt}}\|_F < \epsilon$ **then return** back-pointer path to C
  end if
  if $n \geq N_{\max}$ **then return** NONE
  end if
  $n \leftarrow n + 1$
  for all $C' \in \textsc{Expand}(C, \delta)$ **do** ▷ lattice moves, each $d_J(C, C') \leq \delta$
    $R' \leftarrow \mathrm{F}(C')$; $s \leftarrow \|R - R'\|_F$
    if $R' = \perp \ \vee \ \rho(C') \in \{\emptyset, S\} \ \vee \ s \geq \epsilon$ **then continue**
    end if
    $w \leftarrow \alpha\, d_J(S, \rho(C')) + \beta\, s$
    if $g[S] + w < g[\rho(C')]$ **then**
      $g[\rho(C')] \leftarrow g[S] + w$; set back-pointer $C' \to C$
      push $\big(g[\rho(C')] + h(R'),\, C',\, R'\big)$ onto Q
    end if
  end for
end while
return NONE

function $\mathrm{F}(C)$ ▷ DirectLiNGAM + total effect, memoized on $\rho(C)$
  if $|\rho(C)| < k_{\min}$ **or** fit fails **return** $\perp$
  $A \leftarrow \textsc{DirectLiNGAM}(X[\rho(C)])$
  return $(I - A)^{-1} - I$
end function

The algorithm for computing can be described as follows. First, given a δ-bound on the Jaccard metric on conditions, we construct a lattice starting from C_0 and extending out in all possible directions. Each direction is restricted to a change in the condition C_0 that is linguistically sensible. As an example, given C_0 is molality < 1.46 and monomer_mw < 479 in a dataset with 3 features, then the options for changing the condition in a linguistically reasonable way are: include the third condition with the inequality sign pointing in either direction, move an existing condition by some small value relative to the distribution, or removing an existing condition if the underlying dataset doesn't change by more than δ.

In this construction, we can perform a $A*$ search along the lattice, where edges along the lattice are admissible only if their causal graph distance is less than ϵ. By design their Jaccard distance is less than δ. the $A*$ heuristic for optimal paths can be given by d_{TE}, or a linear combination $\alpha d_J + \beta d_{TE}$. It should be noted that the $A*$ algorithm is guaranteed to admit an optimal path only if the heuristic never overestimates. For this work, $A*$ is only being used for the existence of paths, but future work (8.2) involves optimality. Beam search is another alternative, particularly when one is only interested in the existence of a path, but $A*$ is used in our current implementation.

4 Theoretical Guarantees

4.1 Stability

In the algorithm of DirectLiNGAM [14], Gram matrices with Gaussian kernel values are first used to estimate the dependence, then determine the causal ordering. In this section, we discuss how much it will

change if a single new point is added to the dataset. That is, we give an estimation and error description of the mutual information part of the DirectLiNGAM algorithm.

Let $\{x_i\}_{i=1}^p$ be given columns of a dataset $\mathbf{X}$. That is, each x_i is a random variable, a feature we observed. In the following, we will use y_1, y_2 to denote the two x_i, x_j we are comparing.

Definition 4 (Gaussian Kernel Gram Matrices)**.** Given two sets of n observations of feature y_1 and y_2, the Gaussian kernel Gram matrices are defined as

$$\begin{aligned} K_1(y_1^{(i)}, y_1^{(j)}) &:= \exp(-\frac{1}{2\sigma^2}||y_1^{(i)} - y_1^{(j)}||^2) \\ K_2(y_2^{(i)}, y_2^{(j)}) &:= \exp(-\frac{1}{2\sigma^2}||y_2^{(i)} - y_2^{(j)}||^2), \end{aligned} \tag{1}$$

where $\sigma > 0$ is the bandwidth of Gaussian kernel.

Note that K_1 and K_2 have the same structure but take values from two different random variables. The number of rows and columns of these matrices are both equal to the number of rows on which DirectLiNGAM is run. After this preparation, we can introduce the main tool for estimating the dependence of two random variables.

Definition 5 (Hilbert–Schmidt Independence Criterion (HSIC))**.**

$$HSIC_n(\mathbf{X}, y_1, y_2) := \frac{1}{(n-1)^2} tr(K_1 H_n K_2 H_n), \tag{2}$$

where $H_n := I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$, I_n is the identity matrix and $\mathbf{1}_n$ is the n vector with all entries 1.

The HSIC matrix measures the independence of two random variables.

Definition 6 (Residuals)**.** The residuals when x_i regressed on x_j are defined as

$$r_i^{(j)} = x_i - \frac{cov(x_i, x_j)}{var(x_j)} x_j \tag{3}$$

In [14] lemma 1 and corollary 1 showed that the causal order do not change after we regressed on certain x_j. That is, the residuals derived from regress over x_j, $\{r_i^{(j)}\}$ and the initial random variables $\{x_i\}_{i=1}^p$ have the same causal ordering.

For the equally weighted discrete dataset, let $\bar{x}_i = E[x_i]$

$$\begin{aligned} cov(x_i, x_j) &= E[(x_i - \bar{x}_i)(x_j - \bar{x}_j)] \\ var(x_j) &= cov(x_j, x_j) = E[(x_j - \bar{x}_j)^2] \end{aligned} \tag{4}$$

In our case, the dataset is finite with n rows, and equally weighted for each feature, thus we may write down

$$\begin{aligned} &E[x_i] = \bar{x}_i = \frac{1}{n}\sum_{k=1}^n x_{i,k}, E[x_j] = \bar{x}_j = \frac{1}{n}\sum_{k=1}^n x_{j,k} \\ &\sum_{k=1}^n (x_{i,k} - \bar{x}_i)(x_{j.k} - \bar{x}_j) = n * cov(x_i, x_j) = S_{ij} \end{aligned} \tag{5}$$

and we denote the regression coefficient as

$$\beta := \frac{S_{ij}}{S_{jj}} = \frac{cov(x_i, x_j)}{var(x_j)} \tag{6}$$

Now, we consider how the residuals change if we add a new point in our dataset $\mathbf{X}$. Denote $\tilde{\mathbf{X}} = \mathbf{X} \cup \{p\}$. Write the vector with a new row $\tilde{x_j} = \begin{pmatrix} x_j \\ l_j \end{pmatrix}$ and the new residuals as $\tilde{r}_i^{(j)} = \tilde{x}_i - \tilde{\beta}\tilde{x}_j$.

Let

$$I = l_i - \bar{x}_i, \qquad J = l_j - \bar{x}_j,$$

That is, the difference of value of the added point and the initial mean. Then the updated slope is

$$\tilde{\beta} = \frac{cov(\tilde{x}_i, \tilde{x}_j)}{var(\tilde{x}_j)} = \frac{\sum_{k=1}^{n}(x_{i,k} - \bar{x}_i)(x_{j.k} - \bar{x}_j) + (l_i - \bar{x}_i)(l_j - \bar{x}_j}{\sum_{k=1}^{n}(x_{j.k} - \bar{x}_j)^2 + (l_j - \bar{x_j})^2)} = \frac{S_{ij} + IJ}{S_{jj} + J^2} \tag{7}$$

For our consideration, what we want to estimate is the difference between the dataset we add a new point and the initial dataset.

$$\tilde{\beta} - \beta = \frac{S_{ij} + IJ}{S_{jj} + J^2} - \frac{S_{ij}}{S_{jj}} = \frac{S_{jj}(S_{ij} + IJ) - S_{ij}(S_{jj} + J^2)}{S_{jj}(S_{jj} + J^2)} = \frac{S_{jj}IJ - S_{ij}J^2}{S_{jj}(S_{jj} + J^2)} \tag{8}$$

$$= \frac{J(I - \beta)}{S_{jj} + J^2} \tag{9}$$

Thus we have a bound depends on the new point to average

$$||\tilde{\beta} - \beta|| \leq ||\frac{J(I - \beta)}{S_{jj}}|| = \frac{|J||I - \beta|}{|S_{jj}|} \tag{10}$$

In the algorithm of DirectLiNGAM [14] the method be used to determine the causal order is the mutual method. In the following, we use the HSIC approach to discuss how adding a point will influence the causal ordering, thus the continuity. Notice that the discontinuous points may also tell user information.

Definition 7 (kernel-based independence measure)**.** Let U be the set of indices of features, say $U = \{1, 2, \cdots, p\}$. The kernel-based independence measure is defined as

$$T_{kernel}(x_j, U) = \sum_{i \in U, i \neq j} \hat{MI}_{kernel}(x_j, r_i^{(j)}) \tag{11}$$

In the algorithm, kernel-based independence measure is used to determine the causal ordering and it contributes the discontinuity of total effect metric. In the following, we will try to use HSIC to approach the behavior of adding a point to dataset, how much it may change the causal ordering.

Let $N = n + 1$, look at

$$|\tilde{T}_{kernel}(x_j, U) - T_{kernel}(x_j, U)| = |\sum_{i \in U, i \neq j} HSIC_N(\tilde{\mathbf{X}}, \tilde{x}_j, \tilde{r}_i^{(j)}) - HSIC_n(\mathbf{X}, x_j, r_i^{(j)})| \tag{12}$$

$$= |\sum_{i \in U, i \neq j} \frac{1}{(N-1)^2} tr(\tilde{K}_1 H_N \tilde{K}_2 H_n) - \frac{1}{(n-1)^2} tr(K_1 H_n K_2 H_n)| \tag{13}$$

$$\tag{14}$$

To deal with things inside trace. Let us first discuss the $tr(K_1 H_n K_2 H_2)$. Let $J_n = \mathbb{1}_n \mathbb{1}_n^T$

Lemma 6 (trace calculation formulas)**.** *1.* $tr(ABC) = tr(BCA) = tr(CAB)$

2. $tr(AJ_n) = \mathbf{1}_n^T A \mathbf{1}_n$

Thus we have

$$tr(K_1 H_n K_2 H_2) = tr(K_1(I_n - J_n)K_2(I_n - J_n)) \tag{15}$$

$$= tr(K_1 K_2) - \frac{1}{n} tr(K_1 K_2 J_n) - \frac{1}{n} tr(K_1 J_n K_2) + \frac{1}{n^2}(K_1 J_n K_2 J_n) \tag{16}$$

$$= tr(K_1 K_2) - \frac{2}{n} \mathbf{1}_n^T K_1 K_2 \mathbf{1}_n + \frac{1}{n^2} \mathbf{1}_n^T K_1 \mathbf{1}_n \mathbf{1}_n^T K_2 \mathbf{1}_n \tag{17}$$

For the adding one new point case, we have totally the same form

$$tr(\tilde{K}_1 H_n \tilde{K}_2 H_2) = tr(\tilde{K}_1 (I_n - J_N)\tilde{K}_2 (I_n - J_N)) \tag{18}$$

$$= tr(\tilde{K}_1\tilde{K}_2) - \frac{1}{n}tr(\tilde{K}_1\tilde{K}_2 J_n) - \frac{1}{n}tr(\tilde{K}_1 J_n \tilde{K}_2) + \frac{1}{n^2}(\tilde{K}_1 J_n \tilde{K}_2 J_n) \tag{19}$$

$$= tr(\tilde{K}_1\tilde{K}_2) - \frac{2}{n}\mathbf{1}_n^T\tilde{K}_1\tilde{K}_2\mathbf{1}_n + \frac{1}{n^2}\mathbf{1}_n^T\tilde{K}_1\mathbf{1}_n\mathbf{1}_n^T\tilde{K}_2\mathbf{1}_n \tag{20}$$

Where $\tilde{K}_1$ and $\tilde{K}_2$ can be written as

$$\tilde{K}_1 = \begin{pmatrix} K_1' & k_1 \\ k_1^T & 1 \end{pmatrix}, \qquad \tilde{K}_2 = \begin{pmatrix} K_2' & k_2 \\ k_2^T & 1 \end{pmatrix},$$

and since $\tilde{x}_j = \begin{pmatrix} x_j \\ l_j \end{pmatrix}$, we have

$$(K_1')_{a,b} = \exp\left(-\frac{1}{2\sigma^2}\left\|\tilde{x}_j^{(a)} - \tilde{x}_j^{(b)}\right\|^2\right) = \exp\left(-\frac{1}{2\sigma^2}||x_j^{(a)} - x_j^{(b)}||^2\right) = (K_1)_{a,b} \tag{21}$$

for $a, b \leq n+1$.

k_1 is a n-column vector with $(k_1)^{(a)} = \exp\left(-\frac{1}{2\sigma^2}||x_j^{(a)} - l_j||^2\right)$,

For the $\tilde{K}_2'$ part, the situation is more complicated due to the regression coefficient problem.

$$(K_2')_{a,b} = \exp\left(-\frac{1}{2\sigma^2}\left\|(\tilde{r}_i^{(j)})^{(a)} - (\tilde{r}_i^{(j)})^{(b)}\right\|^2\right) = \exp\left(-\frac{1}{2\sigma^2}\left\|(\tilde{x}_i^{(a)} - \tilde{\beta}\tilde{x}_j^{(a)}) - (\tilde{x}_i^{(b)} - \tilde{\beta}\tilde{x}_j^{(b)})\right\|^2\right) \tag{22}$$

$$= \exp\left(-\frac{1}{2\sigma^2}\left\|(\tilde{x}_i^{(a)} - \tilde{x}_i^{(b)}) - \tilde{\beta}(\tilde{x}_j^{(a)} - \tilde{x}_j^{(b)})\right\|^2\right) = \exp\left(-\frac{1}{2\sigma^2}\left\|(x_i^{(a)} - x_i^{(b)}) - \tilde{\beta}(x_j^{(a)} - x_j^{(b)})\right\|^2\right) \tag{23}$$

for $a, b \leq n+1$.

k_2 is a n-column vector with

$$(k_2)^{(a)} = \exp\left(-\frac{1}{2\sigma^2}\left\|(\tilde{r}_i^{(j)})^{(a)} - (\tilde{r}_i^{(j)})^{(n+1)}\right\|^2\right) = \exp\left(-\frac{1}{2\sigma^2}\left\|(\tilde{x}_i^{(a)} - \tilde{x}_i^{(n+1)}) - \tilde{\beta}(\tilde{x}_j^{(a)} - \tilde{x}_j^{(n+1)})\right\|^2\right) \tag{24}$$

It seems natural to ask the following questions :

1. How much the difference between K_1, K_1' and K_2 , K_2'
2. How much the size of the dataset influences the error

To answer the questions, first we expand the terms inside

$$tr(\tilde{K}_1 H_n \tilde{K}_2 H_2) = tr(\tilde{K}_1\tilde{K}_2) - \frac{2}{n}\mathbf{1}_n^T\tilde{K}_1\tilde{K}_2\mathbf{1}_n + \frac{1}{n^2}\mathbf{1}_n^T\tilde{K}_1\mathbf{1}_n\mathbf{1}_n^T\tilde{K}_2\mathbf{1}_n \tag{25}$$

All the calculation detail of expansion can be found in appendix.

To compare the difference of K_1, K_1', K_2, K_2', we need the following lemma

Lemma 7 (Lipschitz property for Gaussian kernel)**.** *Let $K(x,y) = \exp(\frac{-1}{2\sigma^2}||x-y||^2)$ be the Gaussian kernel, then it satisfies Lipschitz continuous for on variable.*

$$||K(x,y) - K(x,z)|| \leq L||y - z|| \tag{26}$$

Moreover, $L = \frac{1}{\sigma\sqrt{e}}$

Proof :

Let $f(z) = \exp(\frac{-1}{2\sigma^2}||z||^2)$, then $\nabla f(z) = \frac{-z}{\sigma^2}\exp(\frac{-1}{2\sigma^2}||z||^2)$, thus $||\nabla f(z)|| = \frac{||z||}{\sigma^2}\exp(\frac{-1}{2\sigma^2}||z||^2)$. The mean value theorem tells us that

$$|f(x) - f(y)| \leq sup_{\xi}|\nabla f(\xi)||x - y| \tag{27}$$

Thus once we find out a maximum of $|\nabla f(x)|$, we can find out the estimation coefficient we need.

Let $h(r) \coloneqq \frac{r}{\sigma^2}\exp(\frac{-1}{2\sigma^2}r^2)$, then

$$\frac{\partial}{\partial r}h(r) = \frac{-1}{\sigma^2}\exp(\frac{-1}{2\sigma^2}r^2) + \frac{r^2}{\sigma^4}\exp(\frac{-1}{2\sigma^2}r^2) = (1 - \frac{r^2}{\sigma^2})\frac{1}{\sigma^2}\exp(\frac{-1}{2\sigma^2}r^2) \tag{28}$$

When $r = \sigma$, $h(r)$ admits its maximum, so $|\nabla f(z)|$ admits its maximum when $||z|| = \sigma$, it takes the value $\frac{1}{\sigma\sqrt{e}}$

Since $K_1' = K_1$, the only term we need to estimate is the difference between K_2' and K_2. Due to the

$$(K_2')_{a,b} - (K_2)_{a,b} = \exp\left(-\frac{1}{2\sigma^2}\left\|(x_i^{(a)} - x_i^{(b)}) - \tilde{\beta}(x_j^{(a)} - x_j^{(b)})\right\|^2\right) - \exp\left(-\frac{1}{2\sigma^2}\left\|(x_i^{(a)} - x_i^{(b)}) - \beta(x_j^{(a)} - x_j^{(b)})\right\|^2\right) \tag{29}$$

Lemma 8.

$$|\,\|A\|^2 - \|B\|^2\,| = |\,< A + B, A - B >\,| \tag{30}$$

where $< x, y >$ denotes the inner product for two vectors x, y

Let $(x_i^{(a)} - x_i^{(b)}) = u$ and $(x_j^{(a)} - x_j^{(b)}) = v$, and use the Lipschitz property

$$(K_2')_{a,b} - (K_2)_{a,b} = \exp\left(-\frac{1}{2\sigma^2}\left\|u - \tilde{\beta}v\right\|^2\right) - \exp\left(-\frac{1}{2\sigma^2}\|u - \beta v\|^2\right) \tag{31}$$

$$\leq \frac{1}{\sigma\sqrt{e}}|\left\|u - \tilde{\beta}v\right\|^2 - \|u - \beta v\|^2\,| \tag{32}$$

Take $A = u - \tilde{\beta}v$, $B = u - \beta v$, by lemma, we have

$$|\left\|u - \tilde{\beta}v\right\|^2 - \|u - \beta v\|^2\,| = |\,< (u - \tilde{\beta}v) + (u - \beta v), (u - \tilde{\beta}v) - (u - \beta v) >\,| \tag{33}$$

$$= |\,< 2u - (\tilde{\beta} + \beta)v, -(\tilde{\beta} - \beta)v >\,| = |\tilde{\beta} - \beta|\,| < 2u - (\tilde{\beta} + \beta)v, v >\,| \tag{34}$$

$$\leq |\tilde{\beta} - \beta| \cdot ||2u - (\tilde{\beta} + \beta)v|| \cdot ||v|| \tag{35}$$

We may write

$$K_2' = K_2 + \Delta K_2 \tag{36}$$

where $\Delta K_2 = O(|\tilde{\beta} - \beta|)$
Now

$$\left|\tilde{T}_{\text{kernel}}(x_j, U) - T_{\text{kernel}}(x_j, U)\right| = \left|\sum_{i \in U, i \neq j} HSIC_N(\tilde{\mathbf{X}}, \tilde{x}_j, \tilde{r}_i^{(j)}) - HSIC_n(\mathbf{X}, x_j, r_i^{(j)})\right| \tag{37}$$

$$= \left|\sum_{i \in U, i \neq j}\left[\frac{1}{(N-1)^2}\left\{\text{tr}(\tilde{K}_1\tilde{K}_2) - \frac{2}{N}\mathbf{1}_n^T\tilde{K}_1\tilde{K}_2\mathbf{1}_n + \frac{1}{N^2}\mathbf{1}_n^T\tilde{K}_1\mathbf{1}_n\,\mathbf{1}_n^T\tilde{K}_2\mathbf{1}_n\right\}\right.\right. \tag{38}$$

$$\left.\left. - \frac{1}{(n-1)^2}\left\{\text{tr}(K_1K_2) - \frac{2}{n}\mathbf{1}_n^TK_1K_2\mathbf{1}_n + \frac{1}{n^2}\mathbf{1}_n^TK_1\mathbf{1}_n\,\mathbf{1}_n^TK_2\mathbf{1}_n\right\}\right]\right|. \tag{39}$$

Readers can check the calculation details in the appendix.

After expansive calculation on each term, we have

$$= \left| \sum_{i\in U, i\neq j} \left[\frac{1}{n^2}\operatorname{tr}(K_1\Delta K_2) - \frac{2}{n^3}\operatorname{tr}(K_1K_2) - \frac{2}{n^3}\mathbf{1}_n^T K_1 \Delta K_2 \mathbf{1}_n \right. \tag{40}$$

$$- \frac{2}{n^3}\Big(\mathbf{1}_n^T K_1 k_2 + k_1^T K_2 \mathbf{1}_n + k_1^T \Delta K_2 \mathbf{1}_n + k_1^T k_2 + (\mathbf{1}_n^T k_1)(k_2^T \mathbf{1}_n) + \mathbf{1}_n^T k_1 + k_2^T \mathbf{1}_n + 1\Big) \tag{41}$$

$$+ \frac{2}{n^2} k_1^T k_2 + \frac{1}{n^2} \tag{42}$$

$$+ \frac{1}{n^2(n+1)^2}\Big((\mathbf{1}_n^T K_1 \mathbf{1}_n)(\mathbf{1}_n^T \Delta K_2 \mathbf{1}_n) + 2(\mathbf{1}_n^T K_1 \mathbf{1}_n)(k_2^T \mathbf{1}_n) + \mathbf{1}_n^T K_1 \mathbf{1}_n \tag{43}$$

$$+ 2(k_1^T \mathbf{1}_n)(\mathbf{1}_n^T K_2 \mathbf{1}_n) + 2(k_1^T \mathbf{1}_n)(\mathbf{1}_n^T \Delta K_2 \mathbf{1}_n) + 4(k_1^T \mathbf{1}_n)(k_2^T \mathbf{1}_n) \tag{44}$$

$$\left. + 2k_1^T \mathbf{1}_n + \mathbf{1}_n^T K_2 \mathbf{1}_n + \mathbf{1}_n^T \Delta K_2 \mathbf{1}_n + 2k_2^T \mathbf{1}_n + 1\Big) + o(n^{-4}) \right] \Bigg|. \tag{45}$$

Lemma 9 (Cauchy–Schwarz inequality for Frobenius norm and inner product)**.** *Let A,B be two matrices, the Frobenius inner product is defined as*

$$< A, B >_F = tr(A^T B)$$

and it satisfies the Cauchy-Schwarz inequality

$$| < A, B >_F | \leq ||A||_F ||B||_F$$

By this lemma, we have

$$tr(K_1 \Delta K_2) \leq ||K_1||_F ||\Delta K_2||_F = o(J) \tag{46}$$

In conclusion, we have the following estimation

$$|\tilde{T}_{kernel}(x_j, U) - T_{kernel}(x_j, U)| = O(n^{-2}) \tag{47}$$

In the above discussion, we get a error estimation for adding a new point into our dataset, we can understand it as an $o(n^{-2})$ speed. In a more precise way, we can see that almost $o(n^{-3})$ and the $o(n^2)$ term coefficient are actually depends on the distance of the new point added $|J|$.

We still need to discuss whether add a point will change the causal order. Since T_{kernel} is used to determine the causal ordering in the DirectLiNGAM via comparing all the relevance.

In DirectLiNGAM, the causal ordering is determined by comparing the $T_{kernel}(x_i)$ of each feature x_i, that is, we want to find out the

$$x_j = arg\min_{l\in U} T_{kernel}(x_l; U) \tag{48}$$

To make sure the causal ordering do no change under adding a new point. We need to show that $\forall k$,

$$\tilde{T}_{kernel}(x_k) > \tilde{T}_{kernel}(x_j) \tag{49}$$

Here x_i are denotes the features under consideration, inside the matrix $\tilde{x_i}$ denotes the column vector $\begin{pmatrix} x_i \\ l_i \end{pmatrix}$.

Since $x_j = arg\min_{l\in U} T_{kernel}(x_l; U)$, we have $T_{kernel}(x_k) > T_{kernel}(x_j)$ for all k.

If we let $\epsilon_k = T_{\text{kernel}}(x_k) - T_{\text{kernel}}(x_j) > 0$, then the condition becomes

$$\begin{aligned}\forall\, k, \qquad \tilde{T}_{\text{kernel}}(x_k) - \tilde{T}_{\text{kernel}}(x_j) &= \tilde{T}_{\text{kernel}}(x_k) - T_{\text{kernel}}(x_k) \\ &\quad + T_{\text{kernel}}(x_k) - T_{\text{kernel}}(x_j) \\ &\quad + T_{\text{kernel}}(x_j) - \tilde{T}_{\text{kernel}}(x_j) \\ &= \tilde{T}_{\text{kernel}}(x_k) - T_{\text{kernel}}(x_k) + \epsilon_k + T_{\text{kernel}}(x_j) - \tilde{T}_{\text{kernel}}(x_j) \\ &\geq \epsilon_k + (\tilde{T}_{\text{kernel}}(x_k) - T_{\text{kernel}}(x_k)) + (T_{\text{kernel}}(x_j) - \tilde{T}_{\text{kernel}}(x_j)) > 0. \end{aligned} \tag{50}$$

Thus, the causal order remains the same if

Proposition 10 (The causal ordering stable requirement)**.**

$$\epsilon_k > |(\tilde{T}_{\text{kernel}}(x_k) - T_{\text{kernel}}(x_k)) + (T_{\text{kernel}}(x_j) - \tilde{T}_{\text{kernel}}(x_j))| \tag{51}$$

5 Topology of Causal Graph Space

5.1 Why Study Topology?

As materials science experiments are often run under a variety of conditions, each row in a large materials science dataset can be represented in $\mathbb{R}^m$ for an m-column dataset. If many data points are taken under one set of physical conditions, one would expect them to lie in a cluster in $\mathbb{R}^m$. Interesting topological features of the data begin to arise depending on where these clusters of like-condition data are embedded in $\mathbb{R}^m$, the density of the clusters, and whether they form more discrete point clouds or intermingle. In practice, data scientists use topological data analysis (TDA) techniques such as considering the persistent homology of a filtration induced by an underlying Vietoris-Rips complex. Rather than the simple embedding into $\mathbb{R}^m$, one can also look at other topological embeddings of the data and their corresponding Čech complexes, including manifolds with boundary. [16] In this way, filtrations corresponding to large r-balls will have homology approaching the embedded space. The "starting point" of these barcodes may be helpful for determining how "tightly" the data wraps around holes in the embedded space.

In the next subsection, we look at the topology of the causal graph space as a subset of $M_n(\mathbb{R})$, in pursuit of understanding how the data fills in this space. Even if TDA is not possible, understanding the *winding* of data may indicate causal regime flexibility/rigidity. Specifically, we are interested in measuring how much δ-Jaccard path-connected components wrap around holes generated by cycles in the data.

5.2 The Topology of the Causal Graph Space

Representing graphs by their weighted adjacency matrices, it then becomes of interest to study the topology of

$$\begin{aligned}\text{DAG} &= \{M \in M_n(\mathbb{R}) | \forall k \in \mathbb{N}, \forall i,j \in [n], (M^k)_{i,i} = 0, M_{i,j} \neq 0 \implies M_{j,i} = 0\} \\ &\subseteq DG = \{M \in M_n(\mathbb{R}) | \forall i,j \in [n], M_{i,i} = 0, M_{i,j} \neq 0 \implies M_{j,i} = 0\},\end{aligned}$$

the directed acyclic graph space as a subset of the directed graph space (with no self-nodes), in the ℓ_2-norm (Frobenius norm). As the total effect operation on matrices is a homeomorphism, results here are ported over to results on total-effect matrices too. Notice that DAG deformation retracts onto the 0-matrix via $H(M,t) = (1-t)M$, so the more interesting object to study is $\tilde{\text{DAG}} = \text{DAG}\backslash\{0\}$. Intuitively, this allows a deformation retract of $\tilde{\text{DAG}}$ onto Seidel matrices (signed adjacency matrices) so that we may study causal graphs with edges that have significant weight:

$$H(M,t)_{ij} = \begin{cases} (1-t)M_{ij} + t, & M_{ij} > 0, \\ (1-t)M_{ij} - t, & M_{ij} < 0, \\ 0, & M_{ij} = 0. \end{cases}$$

As each DAG with m edges gives an orthant $(0,\infty)^m$, then the topology of $\tilde{\mathrm{DAG}}$ is determined by the poset $(\mathrm{DAG}, \subseteq)$ ordered by edge inclusion. Specifically, $DG_n \simeq (\mathbb{R} \vee \mathbb{R})^{\binom{n}{2}}$, so $\tilde{\mathrm{DAG}} = \bigsqcup_{G \in \mathrm{DAG}} (\mathbb{R}^{\times})^{\mathrm{E(G)}}$ with appropriate gluings. Thus,

$$H_k(\mathrm{DG}, \tilde{\mathrm{DAG}}) = \begin{cases} \mathbb{Z}^{N-S}, & k = 0, \\ 0, & k > 0 \end{cases}$$

for $N = 5^{\binom{n}{2}}$ the number of signed directed graphs and $S = \sum_{G \in \mathrm{DAG}} 2^{|E(G)|}$ the number of signed DAGs. Although this stratified topology could admit interesting structure when viewed through the lens of stratified Morse theory, we choose instead to look at a simpler object: $L_n = S^{n(n-1)-1} \cap \mathrm{DAG}_n$, the link of the DAG space over $S^{n(n-1)-1} = \{M : ||M||_F = 1\}$. As every $\tilde{x} \in \tilde{\mathrm{DAG}}$ admits a projection onto $S^{n(n-1)-1}$ by normalization, this is a natural space to consider for TDA purposes.

Example 4. Consider $L_3 = S^5 \cap \mathrm{DAG}_3$. As subgraph DAGs with $m < \binom{n}{2}$ edges have strata that take the form $L_G \cong \left\{(w_1, \ldots, w_m) \in \mathbb{R}^m_{>0} : \sum_{i=1}^m w_i^2 = 1\right\} = S^{m-1}_+$. Letting C_i be the collection of DAG i-cells, L_3 is composed by gluing C_0, C_1, C_2 along their boundaries. The chains formed from C_i are thus free products of $\mathbb{Z}$ with degrees $(12, 48, 48)$, corresponding to $2 \cdot 3$ choices of a directed edge with 2 sign choices $(2 \cdot 3 \cdot 2 = 12)$, $2^2 \cdot 3$ choices of 2 directed edges with 2^2 sign choices $(2^2 \cdot 3 \cdot 2^2 = 48)$ and $2^3 - 2 = 6$ choices of 3 directed edges with 2^3 sign choices $(6 \cdot 2^3 = 48)$. We may make an exact sequence on these chain complexes using the following definition of a differential. If $\sigma = \big((e_0, \epsilon_0), \ldots, (e_n, \epsilon_n)\big)$ is a signed DAG with $n+1$ edges, ordered lexicographically by the underlying directed edges, then $\partial_n(\sigma) = \sum_{i=0}^n (-1)^i \big((e_0, \epsilon_0), \ldots, \widehat{(e_i, \epsilon_i)}, \ldots, (e_n, \epsilon_n)\big)$. As this is the canonical differential, $\partial_n \circ \partial_{n-1} = 0$. Likewise, DAGs remove an edge produce DAGs, so each differential is well-defined. It suffices to compute $H_k(L_3) = \frac{\mathrm{Ker}(\partial_k)}{\mathrm{Im}(\partial_{k+1})}$ using the rank-Betti formula $\beta_k(L_3) = \dim C_k - \mathrm{rank}(\partial_k) - \mathrm{rank}(\partial_{k+1})$ on finite simplicial complexes. In this instance, we have

$$d_2 = \left(\begin{array}{cccc|c} 1 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 1 & \cdots \\ -1 & 0 & -1 & 0 & \cdots \\ 0 & -1 & 0 & -1 & \cdots \\ \hline \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right)_{12 \times 48}$$

for some canonical ordering and computing the ranks of each d_i in Smith normal form gives the homology:

$$\begin{aligned} \beta_0(L_3) &= \dim C_0 - \mathrm{rank}(\partial_1) = 12 - 11 = 1, \\ \beta_1(L_3) &= \dim C_1 - \mathrm{rank}(\partial_1) - \mathrm{rank}(\partial_2) = 48 - 11 - 36 = 1, \\ \beta_2(L_3) &= \dim C_2 - \mathrm{rank}(\partial_2) - \mathrm{rank}(\partial_3) = 48 - 36 - 0 = 12, \\ \beta_k(L_3) &= 0, \qquad k \geq 3. \end{aligned}$$

$$\boxed{\begin{aligned} &H_0(L_3; \mathbb{Z}) \cong \mathbb{Z}, \\ &H_1(L_3; \mathbb{Z}) \cong \mathbb{Z}, \\ &H_2(L_3; \mathbb{Z}) \cong \mathbb{Z}^{12}, \\ &H_k(L_3; \mathbb{Z}) \cong 0, \qquad k \geq 3. \end{aligned}}$$

The homology of interest in this example is H_1, which is a feature of the "holes" in L_3 that are formed from the clockwise and counterclockwise directed triangle graphs. In practice, this corresponds to an acyclic cycle embedding that cannot be contracted in the space. From a materials science perspective, if data lies near the image of many parameters of this 1-dimensional embedding, it may indicate a contradiction to ground truth or causal regime flexibility. In particular, we have a generator for such an H_1 that corresponds to projecting lines between the following matrices onto S^5 in cyclic order:

$$M_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}, M_2 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}, M_3 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, M_4 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, M_5 = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, M_6 = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

This cycle corresponds to two flips of each arrow in the three-node graph so that cycles are avoided. Looking at data representing this generator in the context of TDA is particularly interesting. Points on "opposite" sides of the loop (i.e. pairs (M_i, M_{i+3})) have causal arrows pointing in the opposite directions, which represent very different causal chains. If the data lies on a cloud for a single condition regime, we would expect only a partial clustering around this generator.

6 Experiments

6.1 Choosing a dataset

To apply Causal AI, we utilized an aggregated tabular dataset in CSV format retrieved from the HTP-MD database (maintained by Toyota Research Institute and MIT at https://www.htpmd.matr.io/)[17]. The dataset comprises 6,270 samples across 11 key features, targeting amorphous polymer-electrolyte systems incorporating LiTFSI salt.The 11 features encompass both structural/simulation attributes and computed transport properties:

- **Structural & Simulation Attributes (6 features):** Trajectory ID, Mol smiles (monomer structure), Molality (salt concentration), Monomer Molecular Weight, Degree of Polymerization, and Density (physical property extracted from MD simulations).
- **Calculated Transport Properties (5 features):** Conductivity (Li^+-ion conductivity), Transference Number, and individual diffusion coefficients for Li^+ (Li^+ Diffusivity), $TFSI^-$ ($TFSI^-$ Diffusivity), and polymer chains (Polymer Diffusivity).

The HTP-MD dataset was selected to evaluate our proposed framework, which fuses Causal AI with mathematical constraints to enhance the interpretability of causal discovery outputs. From a computational perspective, a key reason for choosing this dataset is that its property extraction process is fully transparent. Target transport properties, such as conductivity and diffusivities, are calculated directly from raw molecular dynamics trajectories using the open-source htp-md library based on explicit physical formulations, including mean squared displacement and Cluster Nernst-Einstein approximations. Because the underlying computational logic for these derived properties is known, the dataset serves as a useful benchmark for examining whether our method produces physically interpretable causal graphs. Furthermore, the dataset covers a sequential process in tabular form, connecting input conditions like molality and monomer structural descriptors to an intermediate state variable such as density and target transport properties. Non-numerical attributes, such as raw SMILES strings, are addressed during data preprocessing.In adopting this dataset, we acknowledge its inherent physical limitations. The 5 ns simulation windows may leave long-term polymer chain relaxation incomplete, and the extracted properties rely on physical approximations. As noted by Xie et al. (2022), standard predictive models trained on these features achieve moderate performance, with R^2 values ranging from 0.5 to 0.6, indicating that the simulated targets themselves carry numerical variance and should not be viewed as error-free ground truth.Nevertheless, because all trajectories were generated under a standardized high-throughput protocol with fixed temperature and force fields, unobserved experimental artifacts and reporting biases are effectively minimized. Ultimately, HTP-MD provides a well-structured computational framework with clear property derivation rules. Applying our hybrid approach to this dataset allows us to evaluate the central goal of our work: whether integrating mathematical constraints with Causal AI effectively improves the interpretability of the discovered causal structures in alignment with domain physics.

Acyclicity (Satisfied): The dataset exhibits a three-tier structure: input formulations → intermediate state variables (density) → target transport properties. Physical feedback loops from output transport properties to input formulations do not exist, ensuring a Directed Acyclic Graph (DAG) structure.

Linearity (Partially Violated): Relationships among polymer structures, ion concentrations, and transport properties contain non-linear dependencies. This violation can reduce the accuracy of standard DirectLiNGAM estimations.

Non-Gaussianity and Near-Deterministic Limits (Partially Satisfied / Identification Limit): DirectLiNGAM requires non-Gaussian residual noise to identify causal directions. Although molecular descriptor distributions

and statistical fluctuations in 5 ns MD trajectories provide non-Gaussian signals, certain target properties—such as ionic conductivity and diffusivities—are related via deterministic physical equations (e.g., the Nernst-Einstein relation). Due to the absence of residual noise in near-deterministic relationships, estimated edge directions between these variables become statistically unreliable. These dependencies are therefore better interpreted as physical couplings rather than directional causal paths.

Non-Confoundedness (Satisfied): All observations were generated using a standardized high-throughput MD protocol with uniform force fields and temperatures. This design minimizes unobserved environmental variations, supporting the assumption of error independence.

Sample Size / Infinite Data Assumption (Satisfied): With 6,057 unique samples and a low feature-to-sample ratio ($P = \text{Columns/Rows}$), the dataset provides sufficient sample size to limit estimation variance.

Section Summary: Applicability and Technical Limitations

While the HTP-MD dataset aligns with DirectLiNGAM prerequisites regarding acyclicity, non-confoundedness, and sample size, two main limitations affect the analysis:

- Non-Linearity and Multicollinearity: Correlations among 2D molecular descriptors and non-linear physical responses violate linear model assumptions, increasing the likelihood of spurious edges among structural variables.
- Orientation Ambiguity in Deterministic Equations: For variables governed by exact physical definitions, negligible residual noise prevents DirectLiNGAM from identifying causal directions.

Consequently, edge directions produced by the algorithm cannot be interpreted directly as physical mechanisms without additional domain constraints or treating near-deterministic relationships as non-directional couplings.

- Data Preprocessing and Formatting

6.2 Experiments for Conditional Pathing

The pathing algorithm was applied on polymer simulation data[17]. The primary experiment was done by computing the path components between the entire output of the Fujitsu Causal AI pipeline, then studying the paths in components. We varied three things: the choice of (ϵ, δ) and sample size. The polymer simulation data had a total of around 6500 rows, and we sampled increasingly from 200 to 4000 by adding rows (i.e. we did not resample). Another experiment we ran involved picking a fixed set of conditions, and running the pathing algorithm on those same conditions for some choices of (ϵ, δ) and changing sample sizes. Here, we only computed 3000 and 4000 samples. It should be stated clearly that thus far, we have had very little time to dedicate to studying the output, so the following results and interpretations are preliminary.

We used these computed path components for two things. First, we studied the path components as a form of clustering, looking for relationships differentiating the larger components. Second, we looked for any patterns in the individual paths, looking for connections and structure there.

First, let us look at the individual paths. While our experiments showed us several different paths at different choices of (ϵ, δ) and sample sizes, our analysis, we did not have time to perform a comprehensive analysis of all of the data. Instead, we visually inspected the paths looking for patterns.

Some of the longer paths that we found were long because in if connecting two condition-graph pairs required *dropping* a literal (i.e. the initial condition had some literal while the ending condition did not have that literal), then there were gradual steps to the end of that literal's distribution, upon which the literal was dropped.

Another behavior we saw was that some conditions seemed more important than others to some causal mechanisms. If, along a path, there were conditions that remained fixed, and there were causal mechanisms that remained fixed, then we called these conditions *load bearing candidates* because these conditions may be the primary driver of the fixed causal mechanism. Conversely, if there were conditions that moved freely along that same path, then we call them *mechanism irrelevant candidates* because their contribution to the condition did not impact the existence or strength of the fixed causal mechanism. These are qualitative relationships between the conditions and the causal mechanisms in the path.

Figure 5 shows an example of both of these behaviors. This path connects two condition-graph pairs found from Fujitsu Causal AI on the polymer simulation dataset with 2500 samples. The path was found by running the code that computes the path components with an ϵ-quantile of 0.02 (ϵ *value* of 0.40855), and a δ-quantile of 0.1(δ *value* of 0.43396). The *quantiles* are quantiles of the distribution of pair-wise causal graph distances for this output of the Fujitsu Causal AI. A ϵ-quantile value of 0.02 means that only 2% of the pair-wise total effect distances between the output graphs fall within the true ϵ-value of 0.43396.

In this path, conductivity and tfsi_diffusivity does not change throughout the entire path. Also, the total effect between density and transference number, tfsi_diffusivity and transference number, and conductivity and transference_number are all stable throughout the path. That suggests that conductivity and tfsi_diffusivity are the load bearing candidates here; they are the potential conditions driving these fixed causal mechanisms.

It's important to note here that the Nernst-Einstein equation relate these variables in a consistent way. These fixed mechanisms *should* be related. However, the relationship here is **not** the same relationship found in the Nernst-Einstein equation, it does not follow that simple law. Conditioning did change the behavior. It's unlikely that this behavior is physics-driven, however, for reasons explained in A.

Conversely, we can note that poly_diffusivity changes freely along this path. This is what we called a mechanism irrelevant candidate. Throughout all of these different choices of conditioning on poly_diffusivity, the fixed mechniasms remained. One potential conclusion is that regardless of how this feature is conditioned, the fixed mechanisms remain. Another potential explanation is that *under the other existing conditions*, poly_diffusivity is mechanisms irrelevant. In other words, it's potentially only irrelevant because of it's relationship to other existing conditions in the path. Further experimentation should be done to see what the behavior truly is.

This example showcases some of the potential *structural* findings that the pathing algorithm can find. Conditioning on endogenous variables as described in A breaks the assumptions of Direct LiNGAM, but the structure found in these paths may still lead to new discoveries. That is to say, while the assumptions are broken, and the guarantees lost, that doesn't mean all interpretation is nonexistent. It means that all interpretations must be viewed with the understanding that the guarantees are gone.

Next, we examine the structural differences among the largest path components.

To interpret the resulting (ϵ, δ) path components, we evaluated the total-effect (path-sum) matrices $R \in \mathbb{R}^{p\times p}$ across 557 fixed conditions at $N = 3000$ and $N = 4000$. Rather than relying on DirectLiNGAM's greedy causal ordering—which is sensitive to near-ties that do not reflect true structural shifts—the path components effectively partition the condition space into distinct transport-coupling regimes.

In path-sum space, component centroids differ primarily according to which transport channel couples with electrical conductivity. Inter-centroid distances exceeded intra-component spread by a factor of 3.8, confirming that these components represent fundamentally distinct causal structures rather than arbitrary partitions. Furthermore, cross-N evaluation demonstrated substantial partition reproducibility (ARI 0.39–0.61 vs. an Erdős–Rényi null of 0.00 ± 0.03), indicating that coarse coupled/decoupled splits capture stable physical signatures across independent (Figure 6).

As a concrete example of this partitioning, we evaluated the total-effect values between conductivity and the diffusivity of each ion species ($\mathrm{Li}_{\mathrm{diffusivity}}$ and $\mathrm{TFSI}_{\mathrm{diffusivity}}$), examining whether these causal path-sum strengths varied across components. When the magnitude of the total effect fell below 0.1—indicating an extremely small influence of conductivity on ion diffusivity—we defined the relationship as decoupled; values of 0.1 or greater were classified as coupled. Comparing the components partitioned by our method revealed distinct coupling behaviors: specifically, we identified components characterized by dual-coupling (where conductivity couples with both $\mathrm{Li}_{\mathrm{diffusivity}}$ and $\mathrm{TFSI}_{\mathrm{diffusivity}}$) alongside components where conductivity is decoupled from both species.

These findings demonstrate that path component analysis successfully classifies large ensembles of output causal graphs according to their underlying structural characteristics. By grouping causal graph populations by structural similarity, this approach makes it considerably easier to track transitions in causal structure across varying physical conditions.

Consistent sensitivity trends and high partition agreement were similarly observed across varying sample sizes $N = 3000$ and $N = 4000$(Figure 7).

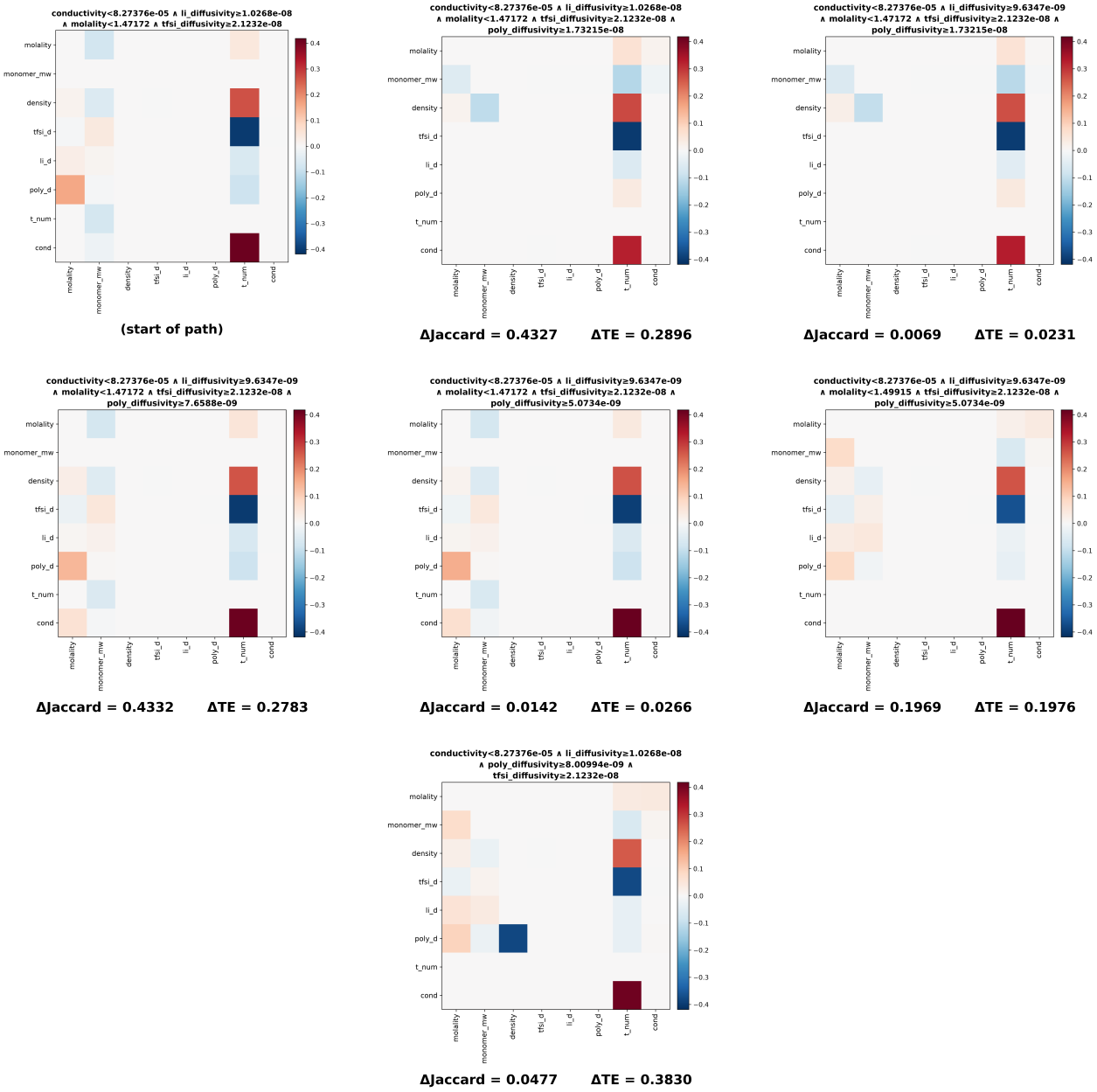


Figure 5: A visual of the path between two condition-graph pairs found from Fujitsu Causal AI. The dataset ran was on the polymer dataset with 2500 samples, an ϵ-quantile of 0.02 (ϵ-*value* of 0.40855), and a δ-quantile of 0.1(δ *value* of 0.43396)

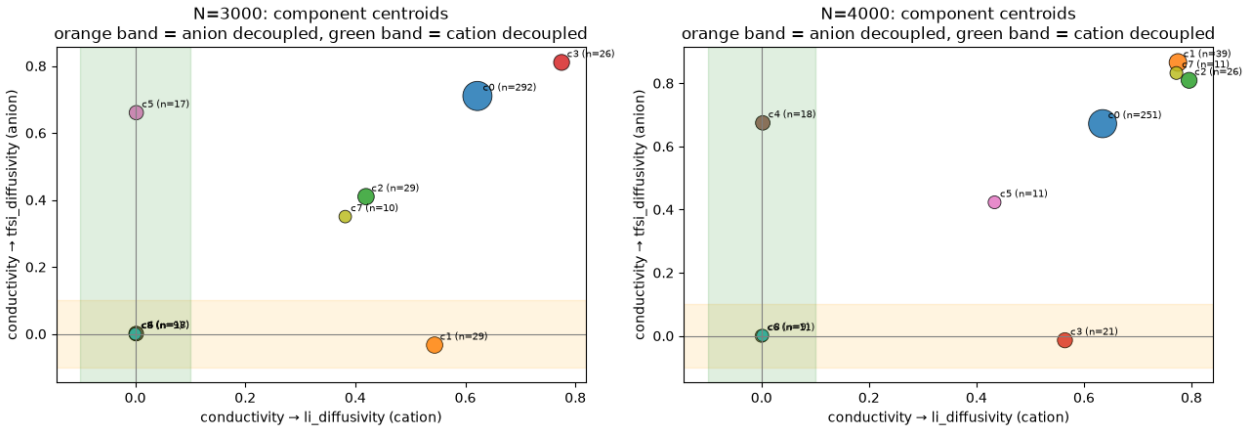


Figure 6: Centroids of the connected components of the inferred causal graph space for sample sizes $N = 3000$ (left) and $N = 4000$ (right), where marker size is proportional to the number of causal dominant graphs in each component, the orange (green) shaded region indicates anion (cation) decoupling from conductivity, and the causal regime remains stable as the sample size increases. The axes represent the decoupling according to cation and anion actions.

6.3 Experiments for Winding Around Cycle Lines

As discussed in the Topology of Causal Graph Space section, we would like to visualize the winding of data about lines that represent CW and CCW cycles on sub-$M_3(\mathbb{R})$ matrices. We consider a projection of the total effect matrices M onto some selected sub-$M_3(\mathbb{R})$:

$$P_{CW} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, P_{CCW} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

for which considering all $\binom{n}{3}$ projections $M|_{P_{CW}}, M|_{P_{CCW}}$ gives no information loss from the original M. Additionally, as there are no 3-cycles in any of these sub-matrices, the projection will lie on the coordinate axes and the coordinate planes in $\mathbb{R}^3$. In light of the L_n link, we normalize all coordinates in each of these projections so that they lie on the sphere. As the data initially lies on the coordinate planes, their normalizations will take them to the 3 great circles formed by the intersections of the coordinate axes with S^2 (see Figure 8). Then, a winding number about this line may be computed by picking a data point, looking at the underlying condition and points near in Jaccard distance up to a threshold, then sequentially choosing the point among these that maximizes CW winding. We place some restrictions on the pathing. Namely, 1) we only choose points such that the CW winding increases, 2) we use a beam search to avoid early termination, 3) we traverse so that every causal sub-order forms a contiguous block on the traversed path (e.x. there cannot be a path $v_i \to v_j \to v_k$ so that v_i, v_k have causal order $(1, 2, 3)$ on the 3 variables and v_j has causal order $(1, 3, 2)$), 4) we avoid pathing so that any edge lies within a cylinder of the cycle line, which limits the winding per step. This cylinder, like the Jaccard distance threshold, is determined by a mean of the value across all data points. Condition 3 prevents "weird" pathing that involves a cut out of these great circles.

structure across condition-graph space that corresponds to a decomposition of regimes into cation and anion conductivity actions. Tuples such as (Molality, Monomer, Density) display more rigid forward winding than reverse winding, which indicates more rigidity in a regime close to the ground truth causal ordering (Molality, Monomer → Density), since the forward longest winding path has points only of ground truth ordering as in Figure 9.

7 Conclusion

The goal of this project was to construct novel methods for analyzing the output of the Fujitsu Causal AI pipeline. We opted for a focus on condition-based movement through the causal graph space. That is to say,

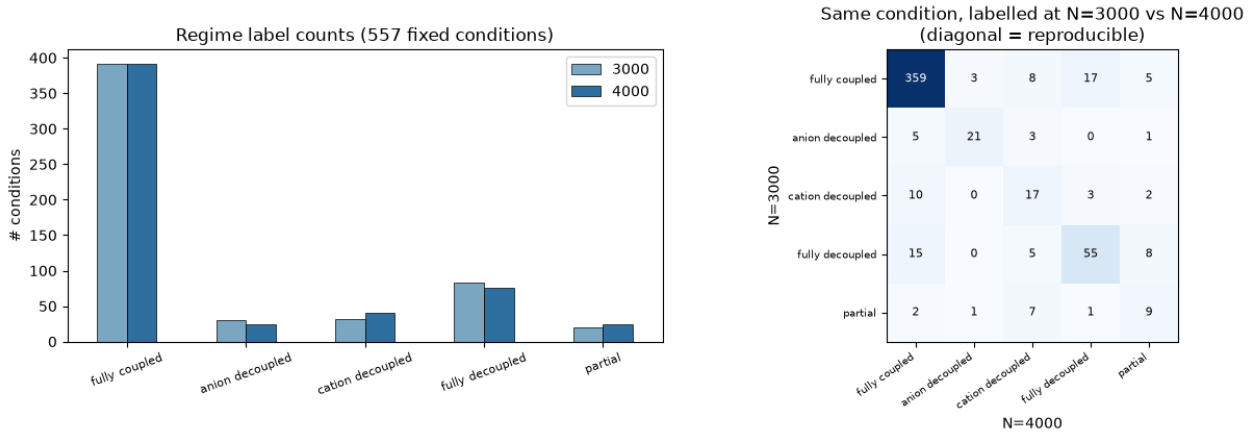


Figure 7: An image showing the robustness of the path component clustering. We sampled at 3000 and 4000 samples and got almost identical path clusters.

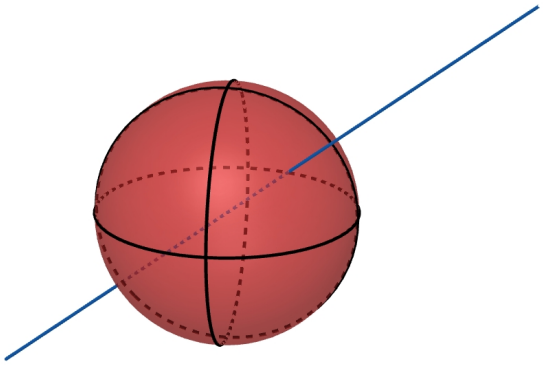


Figure 8: The great circles that data lie on in each projection.

we treated DirectLiNGAM as a function from the space of all possible conditions to the space of all possible causal graphs, and created methods to learn about conditional causality by moving around in these spaces. We ended this project with two primary methods revolving around this idea: conditional pathing and local conditional pathing with winding. Both of these methods aid in **discoverability** by finding condition-based paths, and using those paths to relate certain conditions to certain causal mechanisms.

The conditional pathing algorithm inherently incorporates **comprehensiveness** by design; the condition lattice is as comprehensive as your compute power will allow. δ is a choice of refinement of the discretization of the condition space, and $A*$ and beam search have parameters to allow more or less comprehensiveness.

The winding numbers found from the local pathing approach assign to triples a measure of *flexbility*, describing how much control conditions have over the order of causality in the underlying dataset. Table 1 shows the winding number of all triples in both orientations. It is a **comprehensive** description of the flexbility of the dataset, and the domain scientist may look at these measures and judge for themselves how **convincing** the output of Fujitsu Causal AI is relative to certain substructures by comparing their domain knowledge with the discovered flexibility.

Both of these methods hold promise for improving our understanding conditional causality as a whole, and leave much work to be done. Thus far, the methods have been tried only on small toy data, primarily the commonly used wine dataset [1], and polymer simulation data[17]. Further experiments are required to truly

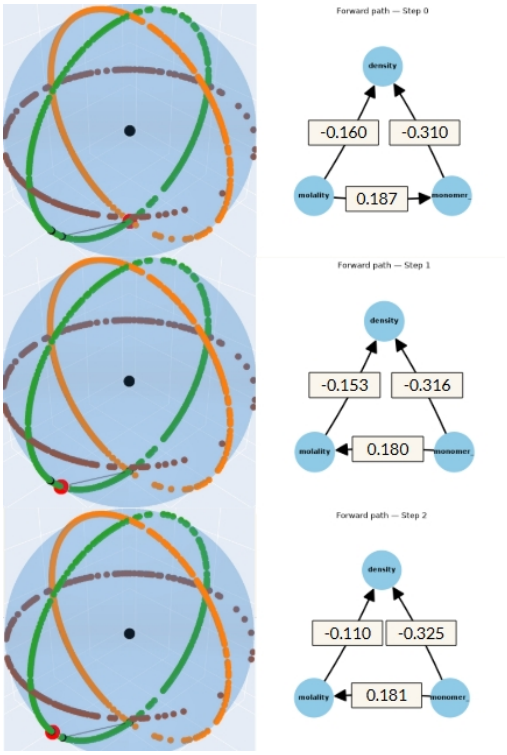


Figure 9: The largest forward winding only has points that match the ground truth in the tuple (Molality, Monomer, Density).

test the capabilities of these new methods. The methods were intentionally designed domain-agnostically, and there may be some types of data that the methods end up being better suited to than others. The methods are inherently topologically and geometrically motivated, so data that is known to showcase nontrivial topology and geometry are good places to start.

Individually, the methods each have future work to consider. The conditional pathing algorithm is slow, and there is much room for optimization in the implementation. Additionally, cleverly choosing the $A*$ heuristic to improve the average time to completion, or even altering this subroutine to keep in mind the awkward geometries of the condition space and causal graph space may go a long way into improving both the results and the efficiency. Finally, The choice of ϵ (the bound in the causal graph movement) has much effect on the output. Researching how to pick a reasonable ϵ, or looking into a solution such as persistence over ϵ, hold potential to greatly improve its utility.

8 Future Work

8.1 Causal Graph Topology Conjectures

Lemma 11. *Let $\widetilde{\Delta}_n^{\mathrm{DAG}}$ denote the signed DAG complex, whose vertices are pairs*

$$(i \to j, \varepsilon), \qquad \varepsilon \in \{-1, +1\},$$

and whose simplices are precisely the collections of signed directed edges whose underlying directed graph is acyclic. Then

$$\beta_{n-2}\left(\widetilde{\Delta}_n^{\mathrm{DAG}}\right) \geq 1,$$

for every $n \geq 3$. In particular,

$$H_{n-2}\left(\widetilde{\Delta}_n^{\mathrm{DAG}}; \mathbb{Z}\right) \neq 0.$$

Proof. Let Δ_n^{DAG} denote the simplicial complex whose vertices are the directed edges

$$(i \to j), \qquad i \neq j,$$

and whose simplices are the edge sets forming directed acyclic graphs. A theorem of Björner and Welker [3] states that $\Delta_n^{\mathrm{DAG}} \simeq S^{n-2}$. Hence, $H_{n-2}\big(\Delta_n^{\mathrm{DAG}};\mathbb{Z}\big) \cong \mathbb{Z}$, so that $\beta_{n-2}\big(\Delta_n^{\mathrm{DAG}}\big) = 1$.

Now, define a simplicial map $p : \widetilde{\Delta}_n^{\mathrm{DAG}} \longrightarrow \Delta_n^{\mathrm{DAG}}$ by forgetting the sign of each directed edge: $p(i \to j, \varepsilon) = (i \to j)$. Since removing the signs does not change the underlying directed graph, p is a well-defined simplicial map. Conversely, choosing the positive sign for every directed edge defines a simplicial inclusion

$$s : \Delta_n^{\mathrm{DAG}} \longrightarrow \widetilde{\Delta}_n^{\mathrm{DAG}}, \qquad s(i \to j) = (i \to j, +1).$$

Clearly, $p \circ s = \mathrm{id}_{\Delta_n^{\mathrm{DAG}}}$. Passing to homology, $p_* \circ s_* = \mathrm{id}_{H_k(\Delta_n^{\mathrm{DAG}};\mathbb{Z})}$ for every k. Therefore, s_* is injective, and taking $k = n-2$ gives $\mathbb{Z} \cong H_{n-2}\big(\Delta_n^{\mathrm{DAG}};\mathbb{Z}\big) \hookrightarrow H_{n-2}\Big(\widetilde{\Delta}_n^{\mathrm{DAG}};\mathbb{Z}\Big)$. Consequently, $H_{n-2}\Big(\widetilde{\Delta}_n^{\mathrm{DAG}};\mathbb{Z}\Big) \neq 0$, and $\beta_{n-2}\Big(\widetilde{\Delta}_n^{\mathrm{DAG}}\Big) \geq 1$. □

The above lemma establishes nontrivial homology in the $n-2$ entry of $H_*(\tilde{\Delta}_n^{\mathrm{DAG}})$. The remaining conjectures seek to better define the higher homologies of the space and understand it as a more complicated topological object.

Conjecture 12. *$H_{n-2}(\tilde{\Delta}_n^{DAG}) = \mathbb{Z}$ for $n > 2$.*

Given the computed homologies in small dimension, we formulate the following spectral sequence conjecture.

	β_0	β_1	β_2	β_3
L_1	1			
L_2	4	1		
L_3	1	1	12	
L_4	1	0	1	0

Definition 8. For $3 \leq r \leq n$, let F_rL_n be the subcomplex of L_n generated by all simplices contained in the boundary of a directed cycle of length at most r. Equivalently,

$$F_rL_n = \bigcup_{\substack{C \text{ directed cycle} \\ |C| \leq r}} \partial C,$$

where ∂C denotes the boundary subcomplex of the simplex on the edge set of C.

This defines an increasing filtration of simplicial complexes

$$F_3L_n \subseteq F_4L_n \subseteq \cdots \subseteq F_nL_n = L_n$$

that induce a first-quadrant spectral sequence

$$E^1_{p,q} \cong H_{p+q}(F_pL_n, F_{p-1}L_n) \Longrightarrow H_{p+q}(L_n),$$

whose E^1-page records the homology contributed by the boundaries of directed cycles of each length.

Conjecture 13. *The spectral sequence collapses at the E_2-page, and its unique surviving nontrivial class in positive degree is represented by the boundary of a directed Hamiltonian cycle. Consequently,*

$$H_{n-2}(L_n;\mathbb{Z}) \cong \mathbb{Z}.$$

It remains of interest to check that filtration is indeed a sensible choice to study the topology of L_n. Methods that may make such homology computations easier include the Quillen fiber theorem and linear algebra tools that reduce the computational complexity of finding Smith normal form boundary matrix ranks.

8.2 Optimal Pathing

In our current implementation, the heuristic for the $A*$ algorithm is optimizing over a scaled combination of the TE distance, Jaccard distance, and change in number of literals (clauses in the conditions such as molality or monomer_mw). The default choice is to only optimize over the TE metrics, as that is our codomain space. This heuristic is a standard choice to make, and is a sort of naive choice of picking what path is the most promising or likely. Our major goal for the pathing algorithm is to find paths that respect the complicated topologies of these spaces, and the metrics are blind to these geometries. For that reason, these choices are not always likely to find the paths that connect our graphs. One potential future direction involves changing this heuristic to somehow include information about the geometry to make more informed searches about the path chosen.

Separately, under the assumption that the heuristic never overestimates, $A*$ gives a guaranteed *optimal path*. Currently, "optimal" is relative to the above linear combination (typically the TE distance), but it need not be. We considered the idea of minimizing over the difference in functionals on the causal graph space. As an example, the 2024 G-RIPS students implemented measurements on a single causal graph called the *hierarchical incoherence coefficient* and the *democracy coefficient*[6]. Both can be considered functionals on the causal graph space, and edge weights between condition-graph pairs can be the difference in these values. Then "optimality" is gradual and minimal change along the path. The general thought is to pick an edge function that preserves whatever one cares about. An immediate thing to note is if the beginning and end point are already chosen, then the values of the functional at those points are fixed, the interesting thing will be how they are changed along the chosen, optimal path. It's not immediately clear what this would give you, so experiments would be necessary to see what value this may bring.

8.3 Picking (ϵ, δ) for pathing, filtrations

It was stated in 3.1 that ϵ and δ are parameters that the users pick when searching for an (ϵ, δ) path between two condition-graph pairs or when computing the path components of a set of condition-graph pairs. In more detail, the choice of δ is a choice of resolution in that it is how much we refine the condition space. There is a trade-off here between refining the grid further, and computational cost. Our implementation for the path components has an input variable `max_nodes` which dictates how much searching we do in the condition space before returning the path components found. This is a large limitation for the current algorithm; the connections found are true, but the separations are not reliable because there may be a path between two components and the program ended before we found it.

Therefore δ is a knob for both *resolution* and *comprehensiveness*. A larger δ allows us to search more of the space, increasing how comprehensive the search is, while decreasing the resolution, resulting in less refined grid sizes. There is work that could be done here in improving the algorithm and increasing the quality of its output. Research into the proper choice of heuristic for the $A*$ search algorithm, for example, may be useful. As the related theory develops, that may also give insight into what paths are even possible, which would also reduce the search space. Additionally, there may be existing literature on such resolution problems, and maybe there are solutions from adjacent aglorithms that can be ported to this one.

Choosing an ϵ is another difficult choice. As we conjecture in 8.4, $\mathcal{F}$ is a piece-wise continuous function with some modulus ω that is independent of the sample size m, so our regions of continuity have a modulus. There are some choice(s) of ϵ that are better than others, but apriori it is not known as the problem is data-dependent.

One potential solution to this, and a research problem in it's own right, is to study a filtration $F' : P(\epsilon_1) \hookrightarrow P(\epsilon_2) \dots P(\max \epsilon)$ of path components that grow as you increase ϵ. This is a very common method in topological data analysis, where filtrations of homology groups are the objects of study. One can sweep ϵ as a pararmeter, and study how long certain path components persist before merging. The components that persist the most are the components most worthy of study, as they are the most "separate" from the other components. This theory is rich with information, and much of the interpretation for such persistent components can be ported over from topological data analysis.

Naturally, the issues with topological data analysis are also ported over. Namely, that of computational complexity and sensitivity to outliers. Over the last twenty years there have been huge steps in decreasing the cost of computing persistence diagrams in topological data analysis, and this has lead to an increase in its use in fields such as medicine, material science, biology, and more. It is not immediately clear if or how these

optimizations can be ported over to this context, but such findings have the potential to make the analysis of these objects practical in real data.

The other problem of sensitivity to outliers is also transferred. In TDA, a single well-placed outlier destroys the persistence diagram, leading to a different (potentially uninformative) interpretation of the data. This has not stopped the method from being very useful in the fields mentioned, but it is a problem to be wary of for the applied users.

8.4 The conjectures of modulus continuity

In the analysis of the input dataset and output causal graph, we still have a stability problem: how does the causal graph change if the user adds new data points? How much does the causal graph change? And how continuous is DirectLiNGAM under the Jaccard metric and the total effect metric? The stability analysis provides an indication of how confidence much the user may have, and we try to determine how the continuity behaves under certain restrictions.

Definition 9. A modulus of continuity is a non-decreasing function $\omega : [0, \infty) \to [0, \infty)$ with $\omega(0) = 0$ and $\lim_{t\to 0} \omega(t) = 0$.

Definition 10. A function $f : (X, d_X) \to (Y, d_Y)$ admits ω as a local modulus of continuity at the point $x \in X$ if and only if $\forall x' \in X,\ d_Y(f(x), f(x')) \leq \omega(d_X(x, x'))$.
If there is a constant L such that $\omega(t) = Lt$, we call f Lipschitz.

Now the question becomes: can we find such an ω such that

$$d_{TE}(F(C), F(C')) \leq \omega(d_J(C, C')) \tag{52}$$

Moreover, can we find an ω independent of the dataset size?

Conjecture 14. *Given a probability distribution $\mathcal{P}$ that satisfies the assumptions of DirectLiNGAM, there exists a modulus of continuity function ω such that*

$$d_{TE}(\mathcal{F}(C), \mathcal{F}(C')) \leq \omega(d_J^{\mu}(C, C')) \tag{53}$$

where d_J^{μ} is the Jaccard metric defined under the measure μ.

Conjecture 15. *Given a dataset X of size m that satisfies the assumptions of DirectLiNGAM, there exists a modulus of continuity function ω_m such that*

$$d_{TE}(\mathcal{F}(C), \mathcal{F}(C')) \leq \omega_m(d_J^m(C, C')) \tag{54}$$

Conjecture 16. *The choice of ω is independent of the size of the dataset. That means we can always find an ω that is sufficient for consistency.*

8.5 Formalization of the condition space and data space

Note that the condition and datasets are living in different spaces. Conditions are composed by clauses and used to describe a data subset. Thus, if we fix conditions and add new data points, we can try to observe how much the Jaccard distance and total effect distance change.

The following formalization is critical for UI, especially for material scientists, physicists, and engineers. It is also critical for mathematicians who need rigorous language to describe the condition space using propositional logic. It may help scientists concretely understand the condition regimes in which they work and distill information. A rigorous definition will now be provided. Fear not, readers who are not familiar with such language should not afraid, for we will also explain the details and philosophy behind such choices.

Our model can be formalized as the following:

Definition 11 (The dataset description)**.** For a dataset $X^{(m)}$ with m data points, the quadruple $(X^m, \mathcal{C}^{(m)}, \Omega^{(m)}, \mathbb{P}^{(m)})$

1. $X^{(m)}$ is the dataset under consideration
2. $\mathcal{C}^{(m)}$ is the set contains all clauses, or the conditions. Thus the proposition logic operations works in this space.
3. $\Omega^{(m)}$ is the sigma algebra on $X^{(m)}$
4. $\mathbb{P}^{(m)}$ denotes the probability function on $X^{(m)}$ with equal distribution.

Then, what we want to describe is

$$(X^m, \mathcal{C}^{(m)}, \Omega^{(m)}, \mathbb{P}^{(m)}) \to (X, \mathcal{C}, \Omega, \mathcal{P}) \text{ as } m \to \infty,$$

where the quadruple $(X, \mathcal{C}, \Omega, \mathcal{P})$ denotes a perfect dataset satisfying the assumptions of DirectLiNGAM.

To formalize the framework considered, we introduce the notion of condition space. In causal analysis for materials science, researchers are often interested in understanding how changes in one or more physical quantities influence other properties of a material. Given an observational dataset, they may wish to investigate the behavior of the system under specified conditions, such as restricting certain variables to prescribed ranges or values.

To describe such conditions in a rigorous and systematic manner, we develop a mathematical language for condition spaces. This formalization provides a common framework for representing conditions independently of any particular algorithm or application. Moreover, it establishes a precise foundation that facilitates communication between researchers developing causal discovery algorithms and domain experts in materials science, enabling both theoretical analysis and practical implementation.

Definition 12 (Condition space)**.** The condition space $\mathcal{C}$ in our discussion can be written as

$$\mathcal{C} = (\mathbb{R} \cup \{None\})^p \times \{<, >, \leq, \geq, None\}^p \tag{55}$$

To explain this definition, we should look back the features we are considering. That is, if we have p features, then given a vector $t \in \mathbb{R}^p$ and a p-tuple with elements $>$ or $<$, then we have a certain condition to cut the $\mathbb{R}^p$. As a simple example, we have have two features, $x_1 =$ density, $x_2 =$ temperature for a certain data set. Then $(x_1 < 3, x_2 > 60)$ is a condition, and we can write it as $t \times s = (3, 60) \times (<, >)$ where $t = (3, 60) \in \mathbb{R}^2$ and $s = (<, >) \in \{<, >, \leq, \geq\}^2$. This vector form actually denotes the condition sentence " We want the elements satisfying the density is lower than 3 and the temperature higher than 60. Readers may note that some times, our restriction could not run out all of the features. The notation "$None$" is thus introduced. For example, if we have three features (x_1, x_2, x_3) and we do not give any restriction on the x_2, we may write $(2, None, 3) \times (\leq, None, >)$ to denote the condition $x_1 < 2$ and $x_3 > 3$.
The above construction has another good property : rather than choose a point t, we can choose a function to describe the boundary condition of the condition chosen.
Until now, the model we have built is simple. Readers who familiar with model theory may realize that there's other approach to make the logic operations on condition space simpler and clearly. To make is readable for experts for different fields including material science, mathematics, algorithms and engineering is one of the future goal. This is the essence idea for user experience design and the scientific discoverability.

The following unsolved problems are mutually motivated and motivating to the formalization language

1. The probability perfectly satisfies DirectLiNGAM and how much the violation of the assumption is acceptable, the stability and the convincingness problem.
2. Whether the finite dataset theory consistent to the perfect dataset with infinite data points.
3. The heterogeneous problem of the dataset, especially for the Jacarrd metric, the gap between the condition space and the embedding of the dataset into Euclidean space. This problem is highly related to the scientific discoverability since it may provide a deep understanding of the physical phenomena.

4. The cutting problem of dataset : how to choose the cutting point, the threshold while we are generating the conditions. Whether there's a good method or certain algorithm and numerical estimation to find out a good threshold in generating the lattice of conditions. To be precisely, can we find out the cutting point or interface which make the phase change or certain physical property totally different?

5. For a perfect dataset, what is a good measure we need. The Lebesgue measure should be used simultaneously with the probability measure to describe the density of dataset. The important part will be the variation of density for discoverability.

6. For a finite dataset, how good the TDA tool, especially the persistence may provide us a measure to done the continuity detection.

7. For the infinite dataset which satisfies the assumption of DirectLiNGAm perfectly, what kind of measure should we use ? Is there exist a natural way to define a universal measure to make it consistent to finite dataset.

A Caveats to Conditional Direct LiNGAM

Wide Learning is one step in the Fujitsu Causal AI algorithm that computes a collection of conditions to maximize some target variable, subject to certain constraints. The conditions are then used to crop the current dataset to a subset that satisfies the conditions, upon which direct LiNGAM is run. Wide learning takes a comprehensive, exhaustive approach to find interesting conditions to apply to direct LiNGAM. This automated workflow seems designed with discovery in mind; it automatically finds new potential causal mechanisms under certain constraints, pointing domain-experts at new potential science. [7]

One danger to this pipeline that may affect the quality of output (and therefore the quality of our above method, should it be used in conjunction with this pipeline), is that the direct LiNGAM as an algorithm is currently blind to the conditions fed into it through wide learning. That is to say, the conditions found by wide learning may break the assumptions that direct LiNGAM requires for quality output. In particular, conditioning on endogenous variables, or variables that are caused by others in the data, creates a *dependence relationship* between the caused and the driver. If $x_i = \alpha x_j + \epsilon$, then x_i is caused by x_j. If we then condition on x_i by requiring $x_i > t$, then we have

$$\begin{aligned} x_i = \alpha x_j + \epsilon &> t \\ \epsilon &> t - \alpha x_j \end{aligned}$$

As the inequality shows the dependence of noise ϵ on x_j, it violates the independence assumption of the directLiNGAM. We in fact see this behavior in our experiments (see 6.2).

Fujitsu's G-RIPS 2025 group studied the assumptions of LiNGAM and the severity of quality control as a function of how far off data is from the assumptions, so there is potential to convert their methodology into an intermediary between wide learning and direct LiNGAM to recover the most robust condition-graph pairs. This is a solution worth investigating, and may further improve the causal discovery pipeline. [11]

B Calculation details in theoretical guarantee

First let us expand the

$$tr(\tilde{K}_1\tilde{K}_2) = tr(\begin{pmatrix} K_1' & k_1 \\ k_1^T & 1 \end{pmatrix}\begin{pmatrix} K_2' & k_2 \\ k_2^T & 1 \end{pmatrix}) \tag{56}$$

$$= tr(\begin{pmatrix} K_1'K_2' + k_1k_2^T & K_1k_2 + k_1 \\ k_1^TK_2 + k_2^T & k_1^Tk_2 + 1 \end{pmatrix}) = tr(K_1'K_2') + 2k_1^Tk_2 + 1 \tag{57}$$

$$\tilde{K}_1\mathbf{1}_N = \begin{pmatrix} K_1' & k_1 \\ k_1^T & 1 \end{pmatrix}\begin{pmatrix} \mathbf{1}_n \\ 1 \end{pmatrix} = \begin{pmatrix} K_1'\mathbf{1}_n + k_1 \\ k_1^T\mathbf{1}_n + 1 \end{pmatrix}, \tilde{K}_2\mathbf{1}_N = \begin{pmatrix} K_2' & k_2 \\ k_2^T & 1 \end{pmatrix}\begin{pmatrix} \mathbf{1}_n \\ 1 \end{pmatrix} = \begin{pmatrix} K_2'\mathbf{1}_n + k_2 \\ k_2^T\mathbf{1}_n + 1 \end{pmatrix} \tag{58}$$

Thus

$$\begin{aligned}
\mathbf{1}_n^T \tilde{K}_1 \tilde{K}_2 \mathbf{1}_n &= (\tilde{K}_1 \mathbf{1}_N)^T \tilde{K}_2 \mathbf{1}_N \\
&= \begin{pmatrix} K_1' \mathbf{1}_n + k_1 \\ k_1^T \mathbf{1}_n + 1 \end{pmatrix}^T \begin{pmatrix} K_2' \mathbf{1}_n + k_2 \\ k_2^T \mathbf{1}_n + 1 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{1}_n^T K_1' + k_1^T & \mathbf{1}_n^T k_1 + 1 \end{pmatrix} \begin{pmatrix} K_2' \mathbf{1}_n + k_2 \\ k_2^T \mathbf{1}_n + 1 \end{pmatrix} \\
&= \mathbf{1}_n^T K_1' K_2' \mathbf{1}_n + \mathbf{1}_n^T K_1' k_2 + k_1^T K_2' \mathbf{1}_n + k_1^T k_2 + \mathbf{1}_n^T k_1 k_2^T \mathbf{1}_n + \mathbf{1}_n^T k_1 + k_2^T \mathbf{1}_n + 1. \qquad (59)
\end{aligned}$$

And

$$\begin{aligned}
\mathbf{1}_N^T \tilde{K}_1 \mathbf{1}_N \, \mathbf{1}_N^T \tilde{K}_2 \mathbf{1}_N &= \begin{pmatrix} \mathbf{1}_n^T & 1 \end{pmatrix} \begin{pmatrix} K_1' & k_1 \\ k_1^T & 1 \end{pmatrix} \begin{pmatrix} \mathbf{1}_n \\ 1 \end{pmatrix} \begin{pmatrix} \mathbf{1}_n^T & 1 \end{pmatrix} \begin{pmatrix} K_2' & k_2 \\ k_2^T & 1 \end{pmatrix} \begin{pmatrix} \mathbf{1}_n \\ 1 \end{pmatrix} \\
&= \left(\mathbf{1}_n^T K_1' \mathbf{1}_n + k_1^T \mathbf{1}_n + \mathbf{1}_n^T k_1 + 1 \right) \left(\mathbf{1}_n^T K_2' \mathbf{1}_n + k_2^T \mathbf{1}_n + \mathbf{1}_n^T k_2 + 1 \right) \\
&= (\mathbf{1}_n^T K_1' \mathbf{1}_n)(\mathbf{1}_n^T K_2' \mathbf{1}_n) + (\mathbf{1}_n^T K_1' \mathbf{1}_n)(k_2^T \mathbf{1}_n) + (\mathbf{1}_n^T K_1' \mathbf{1}_n)(\mathbf{1}_n^T k_2) + \mathbf{1}_n^T K_1' \mathbf{1}_n \\
&\quad + (k_1^T \mathbf{1}_n)(\mathbf{1}_n^T K_2' \mathbf{1}_n) + (k_1^T \mathbf{1}_n)(k_2^T \mathbf{1}_n) + (k_1^T \mathbf{1}_n)(\mathbf{1}_n^T k_2) + k_1^T \mathbf{1}_n \\
&\quad + (\mathbf{1}_n^T k_1)(\mathbf{1}_n^T K_2' \mathbf{1}_n) + (\mathbf{1}_n^T k_1)(k_2^T \mathbf{1}_n) + (\mathbf{1}_n^T k_1)(\mathbf{1}_n^T k_2) + \mathbf{1}_n^T k_1 \\
&\quad + \mathbf{1}_n^T K_2' \mathbf{1}_n + k_2^T \mathbf{1}_n + \mathbf{1}_n^T k_2 + 1. \qquad (60)
\end{aligned}$$

Since $k_i^T \mathbf{1}_n = \mathbf{1}_n^T k_i$ and the fact K_1', K_2' are symmetric, we can make it into the following form

$$\begin{aligned}
\mathbf{1}_N^T \tilde{K}_1 \mathbf{1}_N \, \mathbf{1}_N^T \tilde{K}_2 \mathbf{1}_N &= \left(\mathbf{1}_n^T K_1' \mathbf{1}_n + 2k_1^T \mathbf{1}_n + 1 \right) \left(\mathbf{1}_n^T K_2' \mathbf{1}_n + 2k_2^T \mathbf{1}_n + 1 \right) \\
&= (\mathbf{1}_n^T K_1' \mathbf{1}_n)(\mathbf{1}_n^T K_2' \mathbf{1}_n) + 2(\mathbf{1}_n^T K_1' \mathbf{1}_n)(k_2^T \mathbf{1}_n) + (\mathbf{1}_n^T K_1' \mathbf{1}_n) \\
&\quad + 2(k_1^T \mathbf{1}_n)(\mathbf{1}_n^T K_2' \mathbf{1}_n) + 4(k_1^T \mathbf{1}_n)(k_2^T \mathbf{1}_n) + 2k_1^T \mathbf{1}_n \\
&\quad + \mathbf{1}_n^T K_2' \mathbf{1}_n + 2k_2^T \mathbf{1}_n + 1. \qquad (61)
\end{aligned}$$

For the detail of kernel error estimation, the following

$$\begin{aligned}
\left| \tilde{T}_{\text{kernel}}(x_j, U) - T_{\text{kernel}}(x_j, U) \right| &= \left| \sum_{i \in U, i \neq j} HSIC_N(\tilde{\mathbf{X}}, \tilde{x}_j, \tilde{r}_i^{(j)}) - HSIC_n(\mathbf{X}, x_j, r_i^{(j)}) \right| \qquad (62) \\
&= \left| \sum_{i \in U, i \neq j} \left[\frac{1}{(N-1)^2} \left\{ \operatorname{tr}(\tilde{K}_1 \tilde{K}_2) - \frac{2}{N} \mathbf{1}_n^T \tilde{K}_1 \tilde{K}_2 \mathbf{1}_n + \frac{1}{N^2} \mathbf{1}_n^T \tilde{K}_1 \mathbf{1}_n \, \mathbf{1}_n^T \tilde{K}_2 \mathbf{1}_n \right\} \right. \right. \qquad (63) \\
&\quad \left. \left. - \frac{1}{(n-1)^2} \left\{ \operatorname{tr}(K_1 K_2) - \frac{2}{n} \mathbf{1}_n^T K_1 K_2 \mathbf{1}_n + \frac{1}{n^2} \mathbf{1}_n^T K_1 \mathbf{1}_n \, \mathbf{1}_n^T K_2 \mathbf{1}_n \right\} \right] \right|. \qquad (64)
\end{aligned}$$

and expand it

$$= \Bigg| \sum_{i \in U, i \neq j} \Bigg[\frac{1}{n^2} \Bigg\{ \mathrm{tr}(K_1' K_2') + 2k_1^T k_2 + 1 \tag{65}$$

$$- \frac{2}{n+1} \Big(\mathbf{1}_n^T K_1' K_2' \mathbf{1}_n + \mathbf{1}_n^T K_1' k_2 + k_1^T K_2' \mathbf{1}_n + k_1^T k_2 \tag{66}$$

$$+ \mathbf{1}_n^T k_1 k_2^T \mathbf{1}_n + \mathbf{1}_n^T k_1 + k_2^T \mathbf{1}_n + 1 \Big) \tag{67}$$

$$+ \frac{1}{(n+1)^2} \left(\mathbf{1}_n^T K_1' \mathbf{1}_n + 2k_1^T \mathbf{1}_n + 1 \right) \left(\mathbf{1}_n^T K_2' \mathbf{1}_n + 2k_2^T \mathbf{1}_n + 1 \right) \Bigg\} \Bigg] \tag{68}$$

$$- \frac{1}{(n-1)^2} \left\{ \mathrm{tr}(K_1 K_2) - \frac{2}{n} \mathbf{1}_n^T K_1 K_2 \mathbf{1}_n + \frac{1}{n^2} \mathbf{1}_n^T K_1 \mathbf{1}_n \, \mathbf{1}_n^T K_2 \mathbf{1}_n \right\} \Bigg] \Bigg|. \tag{69}$$

Note that $K_1' = K_1$ and $K_2' = K_2 + \Delta K_2$

$$= \Bigg| \sum_{i \in U, i \neq j} \Bigg[\frac{1}{n^2} \Bigg\{ \mathrm{tr}(K_1 (K_2 + \Delta K_2)) + 2k_1^T k_2 + 1 \tag{70}$$

$$- \frac{2}{n+1} \Big(\mathbf{1}_n^T K_1 (K_2 + \Delta K_2) \mathbf{1}_n + \mathbf{1}_n^T K_1 k_2 + k_1^T (K_2 + \Delta K_2) \mathbf{1}_n + k_1^T k_2 \tag{71}$$

$$+ \mathbf{1}_n^T k_1 k_2^T \mathbf{1}_n + \mathbf{1}_n^T k_1 + k_2^T \mathbf{1}_n + 1 \Big) \tag{72}$$

$$+ \frac{1}{(n+1)^2} \left(\mathbf{1}_n^T K_1 \mathbf{1}_n + 2k_1^T \mathbf{1}_n + 1 \right) \left(\mathbf{1}_n^T (K_2 + \Delta K_2) \mathbf{1}_n + 2k_2^T \mathbf{1}_n + 1 \right) \Bigg\} \tag{73}$$

$$- \frac{1}{(n-1)^2} \left\{ \mathrm{tr}(K_1 K_2) - \frac{2}{n} \mathbf{1}_n^T K_1 K_2 \mathbf{1}_n + \frac{1}{n^2} \mathbf{1}_n^T K_1 \mathbf{1}_n \, \mathbf{1}_n^T K_2 \mathbf{1}_n \right\} \Bigg] \Bigg|. \tag{74}$$

Expand all linear terms we get

$$= \Bigg| \sum_{i \in U, i \neq j} \Bigg[\frac{1}{n^2} \Bigg\{ \mathrm{tr}(K_1 K_2) + \mathrm{tr}(K_1 \Delta K_2) + 2k_1^T k_2 + 1 \tag{75}$$

$$- \frac{2}{n+1} \Big(\mathbf{1}_n^T K_1 K_2 \mathbf{1}_n + \mathbf{1}_n^T K_1 \Delta K_2 \mathbf{1}_n + \mathbf{1}_n^T K_1 k_2 \tag{76}$$

$$+ k_1^T K_2 \mathbf{1}_n + k_1^T \Delta K_2 \mathbf{1}_n + k_1^T k_2 \tag{77}$$

$$+ \mathbf{1}_n^T k_1 k_2^T \mathbf{1}_n + \mathbf{1}_n^T k_1 + k_2^T \mathbf{1}_n + 1 \Big) \tag{78}$$

$$+ \frac{1}{(n+1)^2} \left(\mathbf{1}_n^T K_1 \mathbf{1}_n + 2k_1^T \mathbf{1}_n + 1 \right) \tag{79}$$

$$\times \left(\mathbf{1}_n^T K_2 \mathbf{1}_n + \mathbf{1}_n^T \Delta K_2 \mathbf{1}_n + 2k_2^T \mathbf{1}_n + 1 \right) \Bigg\} \tag{80}$$

$$- \frac{1}{(n-1)^2} \left\{ \mathrm{tr}(K_1 K_2) - \frac{2}{n} \mathbf{1}_n^T K_1 K_2 \mathbf{1}_n + \frac{1}{n^2} \mathbf{1}_n^T K_1 \mathbf{1}_n \, \mathbf{1}_n^T K_2 \mathbf{1}_n \right\} \Bigg] \Bigg|. \tag{81}$$

$$= \Bigg| \sum_{i\in U, i\neq j} \Bigg[\frac{1}{n^2}\Big(\mathrm{tr}(K_1K_2) + \mathrm{tr}(K_1\Delta K_2) + 2k_1^Tk_2 + 1\Big) \tag{82}$$

$$- \frac{2}{n^2(n+1)}\Big(\mathbf{1}_n^TK_1K_2\mathbf{1}_n + \mathbf{1}_n^TK_1\Delta K_2\mathbf{1}_n + \mathbf{1}_n^TK_1k_2 + k_1^TK_2\mathbf{1}_n \tag{83}$$

$$+ k_1^T\Delta K_2\mathbf{1}_n + k_1^Tk_2 + \mathbf{1}_n^Tk_1k_2^T\mathbf{1}_n + \mathbf{1}_n^Tk_1 + k_2^T\mathbf{1}_n + 1\Big) \tag{84}$$

$$+ \frac{1}{n^2(n+1)^2}\Big(\mathbf{1}_n^TK_1\mathbf{1}_n\,\mathbf{1}_n^TK_2\mathbf{1}_n + \mathbf{1}_n^TK_1\mathbf{1}_n\,\mathbf{1}_n^T\Delta K_2\mathbf{1}_n \tag{85}$$

$$+ 2\mathbf{1}_n^TK_1\mathbf{1}_n\,k_2^T\mathbf{1}_n + \mathbf{1}_n^TK_1\mathbf{1}_n \tag{86}$$

$$+ 2k_1^T\mathbf{1}_n\,\mathbf{1}_n^TK_2\mathbf{1}_n + 2k_1^T\mathbf{1}_n\,\mathbf{1}_n^T\Delta K_2\mathbf{1}_n \tag{87}$$

$$+ 4(k_1^T\mathbf{1}_n)(k_2^T\mathbf{1}_n) + 2k_1^T\mathbf{1}_n \tag{88}$$

$$+ \mathbf{1}_n^TK_2\mathbf{1}_n + \mathbf{1}_n^T\Delta K_2\mathbf{1}_n + 2k_2^T\mathbf{1}_n + 1\Big) \tag{89}$$

$$- \frac{1}{(n-1)^2}\,\mathrm{tr}(K_1K_2) + \frac{2}{n(n-1)^2}\mathbf{1}_n^TK_1K_2\mathbf{1}_n \tag{90}$$

$$- \frac{1}{n^2(n-1)^2}\mathbf{1}_n^TK_1\mathbf{1}_n\,\mathbf{1}_n^TK_2\mathbf{1}_n\Bigg]\Bigg|. \tag{91}$$

$$= \Bigg| \sum_{i\in U, i\neq j} \Bigg[\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right)\mathrm{tr}(K_1K_2) + \frac{1}{n^2}\,\mathrm{tr}(K_1\Delta K_2) \tag{92}$$

$$- \frac{2}{n^2(n+1)}\mathbf{1}_n^TK_1K_2\mathbf{1}_n + \frac{2}{n(n-1)^2}\mathbf{1}_n^TK_1K_2\mathbf{1}_n \tag{93}$$

$$- \frac{1}{n^2(n-1)^2}(\mathbf{1}_n^TK_1\mathbf{1}_n)(\mathbf{1}_n^TK_2\mathbf{1}_n) \tag{94}$$

$$+ \frac{1}{n^2(n+1)^2}C_n + \frac{2}{n^2}k_1^Tk_2 + \frac{1}{n^2} \tag{95}$$

$$- \frac{2}{n^2(n+1)}\Big(\mathbf{1}_n^TK_1\Delta K_2\mathbf{1}_n + \mathbf{1}_n^TK_1k_2 + k_1^TK_2\mathbf{1}_n \tag{96}$$

$$+ k_1^T\Delta K_2\mathbf{1}_n + k_1^Tk_2 + (\mathbf{1}_n^Tk_1)(k_2^T\mathbf{1}_n) \tag{97}$$

$$+ \mathbf{1}_n^Tk_1 + k_2^T\mathbf{1}_n + 1\Big)\Bigg]\Bigg|. \tag{98}$$

Since

$$\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right)\mathrm{tr}(K_1K_2) = \frac{1-2n}{n^2(n-1)^2}\,\mathrm{tr}(K_1K_2).$$

$$-\frac{2}{n^2(n+1)} + \frac{2}{n(n-1)^2} = \frac{2(3n-1)}{n^2(n+1)(n-1)^2}.$$

$$\left(\frac{1}{n^2(n+1)^2} - \frac{1}{n^2(n-1)^2}\right)(\mathbf{1}_n^TK_1\mathbf{1}_n)(\mathbf{1}_n^TK_2\mathbf{1}_n) = -\frac{4}{n(n+1)^2(n-1)^2}(\mathbf{1}_n^TK_1\mathbf{1}_n)(\mathbf{1}_n^TK_2\mathbf{1}_n).$$

we have

$$= \Bigg| \sum_{i \in U, i \neq j} \Bigg[\frac{1-2n}{n^2(n-1)^2} \operatorname{tr}(K_1 K_2) + \frac{1}{n^2} \operatorname{tr}(K_1 \Delta K_2) \tag{99}$$

$$+ \frac{2(3n-1)}{n^2(n+1)(n-1)^2} \mathbf{1}_n^T K_1 K_2 \mathbf{1}_n \tag{100}$$

$$- \frac{4}{n(n+1)^2(n-1)^2} (\mathbf{1}_n^T K_1 \mathbf{1}_n)(\mathbf{1}_n^T K_2 \mathbf{1}_n) \tag{101}$$

$$+ \frac{1}{n^2(n+1)^2} \Big(C_n - (\mathbf{1}_n^T K_1 \mathbf{1}_n)(\mathbf{1}_n^T K_2 \mathbf{1}_n) \Big) \tag{102}$$

$$+ \frac{2}{n^2} k_1^T k_2 + \frac{1}{n^2} \tag{103}$$

$$- \frac{2}{n^2(n+1)} \Big(\mathbf{1}_n^T K_1 \Delta K_2 \mathbf{1}_n + \mathbf{1}_n^T K_1 k_2 + k_1^T K_2 \mathbf{1}_n \tag{104}$$

$$+ k_1^T \Delta K_2 \mathbf{1}_n + k_1^T k_2 + (\mathbf{1}_n^T k_1)(k_2^T \mathbf{1}_n) + \mathbf{1}_n^T k_1 + k_2^T \mathbf{1}_n + 1 \Big) \Bigg] \Bigg|. \tag{105}$$

where

$$C_n = \left(\mathbf{1}_n^T K_1 \mathbf{1}_n + 2 k_1^T \mathbf{1}_n + 1 \right) \left(\mathbf{1}_n^T K_2 \mathbf{1}_n + \mathbf{1}_n^T \Delta K_2 \mathbf{1}_n + 2 k_2^T \mathbf{1}_n + 1 \right) \tag{106}$$

$$= (\mathbf{1}_n^T K_1 \mathbf{1}_n)(\mathbf{1}_n^T K_2 \mathbf{1}_n) + (\mathbf{1}_n^T K_1 \mathbf{1}_n)(\mathbf{1}_n^T \Delta K_2 \mathbf{1}_n) \tag{107}$$

$$+ 2(\mathbf{1}_n^T K_1 \mathbf{1}_n)(k_2^T \mathbf{1}_n) + \mathbf{1}_n^T K_1 \mathbf{1}_n \tag{108}$$

$$+ 2(k_1^T \mathbf{1}_n)(\mathbf{1}_n^T K_2 \mathbf{1}_n) + 2(k_1^T \mathbf{1}_n)(\mathbf{1}_n^T \Delta K_2 \mathbf{1}_n) \tag{109}$$

$$+ 4(k_1^T \mathbf{1}_n)(k_2^T \mathbf{1}_n) + 2 k_1^T \mathbf{1}_n \tag{110}$$

$$+ \mathbf{1}_n^T K_2 \mathbf{1}_n + \mathbf{1}_n^T \Delta K_2 \mathbf{1}_n + 2 k_2^T \mathbf{1}_n + 1. \tag{111}$$

Finally we have

$$= \Bigg| \sum_{i \in U, i \neq j} \Bigg[\frac{1}{n^2} \operatorname{tr}(K_1 \Delta K_2) - \frac{2}{n^3} \operatorname{tr}(K_1 K_2) - \frac{2}{n^3} \mathbf{1}_n^T K_1 \Delta K_2 \mathbf{1}_n \tag{112}$$

$$- \frac{2}{n^3} \Big(\mathbf{1}_n^T K_1 k_2 + k_1^T K_2 \mathbf{1}_n + k_1^T \Delta K_2 \mathbf{1}_n + k_1^T k_2 + (\mathbf{1}_n^T k_1)(k_2^T \mathbf{1}_n) + \mathbf{1}_n^T k_1 + k_2^T \mathbf{1}_n + 1 \Big) \tag{113}$$

$$+ \frac{2}{n^2} k_1^T k_2 + \frac{1}{n^2} \tag{114}$$

$$+ \frac{1}{n^2(n+1)^2} \Big((\mathbf{1}_n^T K_1 \mathbf{1}_n)(\mathbf{1}_n^T \Delta K_2 \mathbf{1}_n) + 2(\mathbf{1}_n^T K_1 \mathbf{1}_n)(k_2^T \mathbf{1}_n) + \mathbf{1}_n^T K_1 \mathbf{1}_n \tag{115}$$

$$+ 2(k_1^T \mathbf{1}_n)(\mathbf{1}_n^T K_2 \mathbf{1}_n) + 2(k_1^T \mathbf{1}_n)(\mathbf{1}_n^T \Delta K_2 \mathbf{1}_n) + 4(k_1^T \mathbf{1}_n)(k_2^T \mathbf{1}_n) \tag{116}$$

$$+ 2 k_1^T \mathbf{1}_n + \mathbf{1}_n^T K_2 \mathbf{1}_n + \mathbf{1}_n^T \Delta K_2 \mathbf{1}_n + 2 k_2^T \mathbf{1}_n + 1 \Big) + o(n^{-4}) \Bigg] \Bigg|. \tag{117}$$

C Contributions

C.1 Misaki Nagashima

Misaki was responsible for data set selection and interpretation. As a biologist, she was most well-equipped for interpreting material science data, and spent a large chunk of the project learning about what material science data sets would be amenable to the Fujitsu Causal AI pipeline and to our analysis of the output. After our methods were developed, She interpreted the outputs and checked them against currently known material science knowledge. She also provide the insight in scientific discoverability, in the part of how to understand

the importance of causal graphs and the information scientist would be interested in, related our theoretic work on stability theory development, making sure the feasibility to scientific application. She contributed mainly on the data set and implementation via programming but also provided deep understanding and viewpoint as a biologist and data scientist in our theoretical development section.

C.2 (Brandin) Farris

Farris was the project manager of our group. He was responsible for the schedule and the project itself. As a skillful applied topologist and data scientist, he was able to both implement his algorithmic concepts and generate appealing visuals for them. His strong knowledge of TDA motivated many of the approaches we pursued. He also designed the conditional pathing idea and algorithm. In the mutual work part, Farris worked with Misaki on developing a dataset that would be ideal to run tests on. He worked with Wu on the stability of causal ordering theory and provided critical visualization for limiting cases.

C.3 Wu TEWEI

Wu was responsible for understanding the Fujitsu Causal AI pipeline in detail. Wu studied the Wide Learning algorithm and the Direct LiNGAM algorithm so that any methods we developed fit reasonably well within the pipeline. Later in the program, Wu was leading the formalization and proving of the conjectures regarding the behavior of the Direct LiNGAM algorithm as a function from condition space to causal graph space. Wu and Jonah also collaborated on many of the proofs regarding Direct LiNGAM, playing an instrumental role in fixing mistakes in proofs so that the final product is formally correct. Wu's deep understanding of the entire pipeline has led to the development of more comprehensive theory.

C.4 Jonah Yoshida

Jonah was responsible for understanding the topological spaces involved in this project and developing novel topological methods. Jonah was the lead in the topological theory, studying the causal graph space and it's structure, and using that structure as motivation for a new method for analyzing the output of the Fujitsu Causal AI pipeline. The projection of causal graphs onto the cycle subspace and subsequent analysis of that projection was lead by Jonah. Jonah was also instrumental in proving bounds involving the tie set and detecting, analytically, when the causal ordering changes. Inspired by Farris' conditional pathing idea, Jonah designed the winding algorithm and implemented it.

References

[1] Stefan Aeberhard and Michele Forina. *Wine*. UCI Machine Learning Repository. 1992. DOI: 10.24432/C5PC7J. URL: https://uci.edu.

[2] Mehmet Emin Aktas, Esra Akbas, and Ahmed El Fatmaoui. "Persistence homology of networks: methods and applications". In: *Applied Network Science* 4.1 (2019), p. 61. DOI: 10.1007/s41109-019-0179-3.

[3] Anders Björner and Volkmar Welker. "Complexes of Directed Graphs". In: *SIAM Journal on Discrete Mathematics* (1999).

[4] Samir Chowdhury and Facundo Mémoli. "Persistent path homology of directed networks". In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2018, pp. 1152–1169. URL: https://dl.acm.org/doi/10.5555/3174304.3175345.

[5] Mihir Dhanakshirur et al. "A Continuous Structural Intervention Distance to Compare Causal Graphs". In: *Communications in Computer and Information Science* (2024).

[6] John Forde et al. *G-RIPS Sendai 2024: Fujitsu Group Final Report*. Tech. rep. Mentors: Fabiana Ferracina, Jorge Gutierrez, Hiroyuki Higuchi. Institute for Pure and Applied Mathematics (IPAM), Aug. 2024.

[7] Yuta Fujishige et al. "Causal Discovery Platform in Heterogeneous Populations by Subpopulation Enumeration". In: *Forum on Information Technology (FIT2023)*. Paper F-051. 2023.

[8] Yuta Fujishige et al. "Causal Discovery Platform in Heterogeneous Populations by Subpopulation Enumeration". In: *Proceedings of the National Academy of Sciences* (2023).

[9] Alexander Grigor'yan et al. *Homologies of path complexes and digraphs*. arXiv preprint arXiv:1207.2834. 2012. arXiv: 1207.2834 [math.CO].

[10] Eberhardt Jahn and Schulman. "Lower Bounds on the Size of Markov Equivalence Classes". In: (2025). DOI: 10.48550/arXiv.2506.20933.

[11] Jamin Kochman et al. *Discoverability with Modern AI*. Fujitsu Group Final Report. Mentors: Molly Noel, Jorge Gutierrez, Hiroyuki Higuchi, Kei Fujita. G-RIPS Sendai 2025, Mathematical Science Center for Co-creative Society (MathCCS), Tohoku University, Aug. 7, 2025. URL: https://www.mccs.tohoku.ac.jp/g-rips/report/2025/pdf/fujitsu_final_report.pdf.

[12] Shuhei Obara et al. "Commissioning of a compact multibend achromat lattice NanoTerasu: A new 3 GeV synchrotron radiation facility". In: *Physical Review Accelerators and Beams* 28.2 (2025), p. 020701. DOI: 10.1103/PhysRevAccelBeams.28.020701. URL: https://doi.org/10.1103/PhysRevAccelBeams.28.020701.

[13] Jonas Peters and Peter Bühlmann. *Structural Intervention Distance (SID) for Evaluating Causal Graphs*. 2014. arXiv: 1306.1043 [stat.ML]. URL: https://arxiv.org/abs/1306.1043.

[14] Shohei Shimizu et al. *DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model*. 2011. arXiv: 1101.2489 [stat.ML]. URL: https://arxiv.org/abs/1101.2489.

[15] Judea Pearl Thomas Verma. "Causal Networks: Semantics and Expressiveness". In: *Machine Intelligence and Pattern Recognition* 9.10 (1990), pp. 69–76. DOI: https://www.sciencedirect.com/science/chapter/bookseries/abs/pii/B9780444886507500111.

[16] Katharine Turner. "The extended persistent homology transform of manifolds with boundary". In: *Journal of Applied and Computational Topology* (2024).

[17] Tian Xie et al. "A cloud platform for sharing and automated analysis of raw data from high throughput polymer MD simulations". In: *APL Machine Learning* 1.4 (2023), p. 046108.

Forward			Reverse		
Columns	Winding	Len	Columns	Winding	Len
(M, Li, TN)	1.653	3	(M, MW, D)	1.756	5
(M, MW, TN)	1.573	5	(M, Li, TN)	1.570	5
(MW, D, Li)	1.545	3	(M, MW, TN)	1.217	3
(D, Li, TN)	1.309	3	(M, D, TN)	1.189	3
(D, TN, Cond)	1.074	2	(Li, Poly, TN)	1.182	2
(M, D, TN)	1.051	2	(D, Li, TN)	1.149	3
(MW, TN, Cond)	1.013	4	(M, MW, Cond)	1.132	2
(M, MW, Cond)	0.989	3	(MW, D, Cond)	1.066	2
(TFSI, Poly, TN)	0.892	3	(M, TFSI, Cond)	1.047	4
(MW, Poly, TN)	0.882	4	(TFSI, Li, TN)	1.047	3
(MW, TFSI, Poly)	0.864	4	(MW, Poly, TN)	1.047	4
(M, D, Cond)	0.810	3	(MW, TFSI, Cond)	1.010	3
(MW, D, TN)	0.794	2	(M, Poly, TN)	0.955	2
(M, D, Poly)	0.761	3	(M, TFSI, TN)	0.937	3
(MW, TFSI, Li)	0.732	3	(M, D, Cond)	0.892	2
(Li, Poly, TN)	0.729	3	(M, TN, Cond)	0.873	3
(D, TFSI, Li)	0.706	3	(MW, D, TN)	0.828	3
(M, MW, Poly)	0.700	3	(M, MW, Poly)	0.809	2
(M, MW, D)	0.673	4	(M, MW, Li)	0.770	3
(MW, Li, TN)	0.671	7	(MW, TFSI, TN)	0.754	4
(M, Li, Poly)	0.613	3	(D, TFSI, Cond)	0.734	4
(TFSI, TN, Cond)	0.609	3	(D, Poly, TN)	0.714	4
(TFSI, Li, TN)	0.605	2	(MW, Li, TN)	0.702	3
(M, TN, Cond)	0.600	2	(M, D, Poly)	0.698	3
(D, Li, Poly)	0.592	4	(D, TFSI, TN)	0.661	3
(M, Poly, Cond)	0.579	2	(MW, Li, Poly)	0.648	3
(D, TFSI, Cond)	0.577	2	(TFSI, Li, Poly)	0.643	2
(Poly, TN, Cond)	0.565	2	(M, Li, Cond)	0.640	4
(D, TFSI, Poly)	0.554	2	(D, Poly, Cond)	0.638	4
(MW, D, Cond)	0.543	2	(M, D, Li)	0.622	2
(TFSI, Li, Poly)	0.534	2	(TFSI, TN, Cond)	0.612	3
(M, D, Li)	0.533	2	(MW, TN, Cond)	0.599	2
(M, Poly, TN)	0.531	2	(D, TFSI, Li)	0.588	3
(MW, D, TFSI)	0.521	3	(D, Li, Poly)	0.577	2
(MW, D, Poly)	0.516	2	(M, MW, TFSI)	0.550	2
(M, MW, Li)	0.505	2	(Poly, TN, Cond)	0.543	4
(TFSI, Poly, Cond)	0.489	2	(MW, D, Li)	0.526	2
(M, Li, Cond)	0.478	2	(M, Li, Poly)	0.519	2
(D, Poly, TN)	0.460	2	(D, TN, Cond)	0.515	2
(M, TFSI, Cond)	0.457	2	(M, Poly, Cond)	0.510	2
(Li, Poly, Cond)	0.422	2	(D, Li, Cond)	0.498	2
(M, TFSI, Poly)	0.410	2	(TFSI, Poly, TN)	0.488	2
(D, Li, Cond)	0.380	2	(TFSI, Li, Cond)	0.471	2
(D, TFSI, TN)	0.374	5	(D, TFSI, Poly)	0.452	2
(M, MW, TFSI)	0.324	2	(M, TFSI, Poly)	0.446	2
(MW, Li, Poly)	0.321	3	(MW, Poly, Cond)	0.424	2
(Li, TN, Cond)	0.312	2	(M, TFSI, Li)	0.348	2
(TFSI, Li, Cond)	0.311	2	(MW, Li, Cond)	0.330	2
(D, Poly, Cond)	0.307	2	(MW, TFSI, Poly)	0.277	2
(MW, Poly, Cond)	0.295	2	(TFSI, Poly, Cond)	0.273	3
(M, TFSI, TN)	0.265	3	(M, D, TFSI)	0.245	2
(M, D, TFSI)	0.252	2	(MW, D, Poly)	0.213	2
(MW, TFSI, TN)	0.250	2	(MW, D, TFSI)	0.163	2
(M, TFSI, Li)	0.211	2	(Li, TN, Cond)	0.124	2
(MW, TFSI, Cond)	0.196	2	(MW, TFSI, Li)	0.111	2
(MW, Li, Cond)	0.169	2	(Li, Poly, Cond)	0.095	2

Table 1: Comparison of winding scores and path lengths for all three-variable feature subsets, with forward and reverse winding (in radians).