Alembic®

# A Summary Review of Alembic and Causal Inference for Marketing

# A Summary Review of Alembic and Causal Inference for Marketing

Version 1.0 Sept 6, 2024

# A Summary of Alembic's Causal Inference Approach

*"Alembic is using AI developed for scientific research applications to predict ROI from marketing. NVIDIA marketing is using Alembic with great success."*
*- Jensen Huang, Founder & CEO*

This document addresses key questions about Alembic's causal inference platform, a new approach different than traditional marketing measurement techniques. As modern marketing grows ever more complex with multi-channel campaigns and non-linear interactions, conventional methods like A/B testing and incrementality often fall short. Alembic leverages advanced techniques, such as temporal-spatial directed graph construction[1], to offer a new way to uncover causal relationships.

This document tackles the concerns an experienced data science or analytics team might have directly, covering topics like validating causal estimates, preventing overfitting, and handling large-scale datasets. We aim to provide clear answers and address the uncertainties around adopting causal inference in marketing, offering a grounded perspective on its potential benefits and challenges.

This document is about causal analysis and is not meant to be exhaustive. It has five major topics it focuses on.

- What is the issue and why is there a need for new methods?
- How is the Alembic method "causality"?
- What is a causal chain and where do they come from?
- How do you deal with all this complex spatiotemporal data?
- What questions have you gotten about the methodology?

For a comparison of common non-causal techniques like Marketing Mix Modeling (MMM), last-touch, and other methods, please contact Alembic for additional documentation.

## On the state of causality and marketing attribution

In today's multi-channel marketing environment, traditional models like A/B testing, randomized controlled trials (RCTs), and basic attribution struggle to keep up with the growing complexity of consumer interactions across digital, social, and offline channels.[2] While these methods were once essential for measuring marketing effectiveness, they fall short when applied to modern campaigns that span numerous touchpoints. As consumers engage with brands through

[1] Del Mondo, G., Rodríguez, M. A., Claramunt, C., Bravo, L., & Thibaud, R. (2013). Modeling consistency of spatio-temporal graphs. Data & Knowledge Engineering, 84, 59-80

[2] Kohavi R, Tang D, Xu Y, Hemkens LG, Ioannidis JPA. Online randomized controlled experiments at scale: lessons and extensions to medicine. Trials. 2020 Feb 7;21(1):150. doi: 10.1186/s13063-020-4084-y. PMID: 32033614; PMCID: PMC7007661.

websites, social media, and in-store experiences, it becomes difficult for these models to capture the interconnected nature of large media ecosystems and consumer journeys.

## Why Do We See Limitations in Traditional Methods?

Research highlights their limitations when dealing with large, dynamic datasets across multiple channels and timeframes.[3] While these methods work well in small-scale, controlled settings, they become impractical for large-scale, ongoing campaigns where numerous external factors can influence outcomes.

### What causes the struggle to use traditional causal methods?

Some of the things that make using traditional causal methods extremely difficult to deploy on a holistic basis within large marketing departments.

#### Multi-Channel Complexity, Interdependencies, and externalities

Modern multi-channel campaigns—such as those involving TV ads, social media, and in-store displays—create complex interdependencies that are difficult for traditional methods to untangle from themselves, and externalities[4]. For instance, a TV ad may raise awareness, leading to social media engagement, ultimately resulting in a purchase. Traditional methods like A/B testing and incrementality struggle to model these growing interdependencies accurately.

#### Geographic and Digital Overlap: Managing Spillover Effects

In today's global environment, geographic boundaries are increasingly blurred as campaigns targeted at specific regions often spill over into other markets through digital channels. Traditional methods are limited in addressing these spillover effects[5], which can significantly distort results and complicate campaign performance evaluation.

#### Temporal Dynamics: Short- and Long-Term Effects

Marketing campaigns can generate both immediate sales spikes and long-term brand-building impacts. It’s crucial to account for both when evaluating effectiveness.[6] For example, while a campaign may deliver immediate results, its actual value may manifest over time as it contributes to brand equity growth.

[3] Velummailum RR, McKibbon C, Brenner DR, Stringer EA, Ekstrom L, Dron L. Data Challenges for Externally Controlled Trials: Viewpoint. J Med Internet Res. 2023 Apr 5;25:e43484. doi: 10.2196/43484. PMID: 37018021; PMCID: PMC10132012.

[4] Hall, Robert E. (1986). Market Structure and Macroeconomic Fluctuations. Brookings Papers on Economic Activity, 17(2), 285–338.

[5] Benjamin-Chung J, Arnold BF, Berger D, Luby SP, Miguel E, Colford JM Jr, Hubbard AE. Spillover effects in epidemiology: parameters, study designs and methodological considerations. Int J Epidemiol. 2018 Feb 1;47(1):332-347. doi: 10.1093/ije/dyx201. PMID: 29106568; PMCID: PMC5837695.

[6] Granholm, A., Alhazzani, W., Derde, L.P.G. et al. Randomised clinical trials in critical care: past, present and future. Intensive Care Med 48, 164–178 (2022). https://doi.org/10.1007/s00134-021-06587-9

## Brand Equity and Historical Influence

For brands with a long history of brand-building, isolating the impact of a single campaign can take time and effort. For example—decades of brand-building efforts mean that the success of a recent campaign cannot be solely attributed to that campaign alone. Understanding the cumulative effect of past marketing efforts is essential for accurate analysis.

## There are practical concerns with traditional methods

*Contrary to frequent claims in the applied literature, randomization does not equalize everything other than the treatment in the treatment and control groups, it does not automatically deliver a precise estimate of the average treatment effect (ATE), and it does not relieve us of the need to think about (observed or unobserved) covariates.*[7]

- **A/B testing and RCTs** have traditionally been valuable tools in controlled settings. A/B testing isolates specific elements, such as which ad performs better, while RCTs offer a more structured approach. However, both methods struggle to account for the cumulative effects of multiple campaigns running across various channels. RCTs, in particular, are too rigid to address the complex, fluid nature of real-world marketing.[8]

- **Incrementality** improves on these models by measuring the true impact, or "lift," each marketing effort generates across multiple channels. This provides marketers with better insights into the contributions of individual tactics, allowing for a more nuanced measurement of effectiveness across platforms.

  However, incrementality still lacks a holistic view, as it doesn't fully account for external factors like earned media, competitor actions, or broader market trends. Additionally, it fails to capture the interactions between touchpoints over time, which is crucial for understanding long-term consumer behavior and overall marketing effectiveness.

- **Alembic's spatio-temporal causal inference** observational settings, allowing it to detect causality without needing randomized control groups, handling selection bias more effectively than incrementality. Instead of tracking interactions over time or space separately like incrementality, it captures the full complexity of consumer behavior within a temporal space, analyzing how information flows and influences decision-making across multiple touchpoints. This makes it better suited for complex environments.

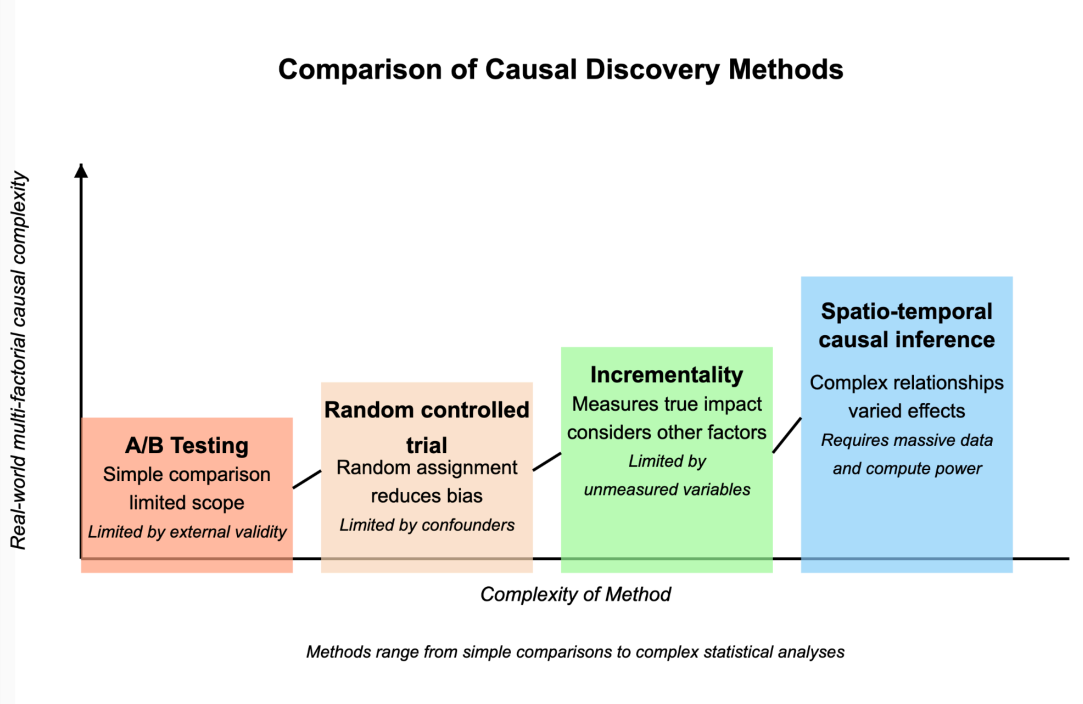
  Our approach also gets around the achilles heel of network and graph analysis, which is

[7] Velummailum RR, McKibbon C, Brenner DR, Stringer EA, Ekstrom L, Dron L. Data Challenges for Externally Controlled Trials: Viewpoint. J Med Internet Res. 2023 Apr 5;25:e43484. doi: 10.2196/43484. PMID: 37018021; PMCID: PMC10132012.

[8] Ellen L. Hamaker, Jeroen D. Mulder, Marinus H. van IJzendoorn, Description, prediction and causation: Methodological challenges of studying child and adolescent development, Developmental Cognitive Neuroscience, Volume 46, 2020, 100867, ISSN 1878-9293,

temporality.[9] It is designed for complex, nonlinear systems, improving the detection of subtle influences in noisy data.

This approach gives marketers a deeper understanding of how campaigns evolve and interact dynamically across platforms, delivering insights that traditional models cannot uncover.[10]



## So, explain why you say this Alembic platform is about causality

*"I find the concept intriguing, but I'm still wrestling with a few points regarding its role in establishing causality. Traditionally, causality has been an exact concept—much more than just a measure of relationships or information flow. It's about identifying a clear, direct link between cause and effect. In classical terms, if I change XXX, I should be able to observe a corresponding change in YYY. The methods we've used over the years—like randomized controlled trials or even Granger causality—are designed with this precision. My concern with Alembic's causal analysis is that while it measures directional information flow, how do we ensure it's not just capturing complex associations but also true causal influence? How do we confirm that the information flow it detects isn't due to some hidden common cause or noise within the system?" - PhD in Data Science & Alembic Customer*

[9] Scholtes, I. "When is a Network a Network? Multi-Order Graphical Model Selection in Pathways and Temporal Networks." KDD'17 - Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Nova Scotia, Canada, August 13-17, 2017. arXiv:1702.05499 [cs.SI].
[10] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11), eaau4996.

**Our answer:**

You're right—traditional causality has always been about direct intervention. Randomized controlled trials allow us to manipulate variables and observe changes, clearly understanding cause and effect. Granger causality, similarly, establishes causality through prediction in time-series data but assumes linearity.

Alembic's causal operates differently, and while it doesn't replace experimental causality, it's a valuable tool in observational settings where direct intervention isn't possible. Alembic's causal measures the directional flow of information from XXX to YYY, accounting for how much knowing XXX's past reduces the uncertainty of YYY's future beyond what YYY's past alone can explain. This conditional aspect is critical—it's not simply an association but a measure of how XXX actively informs YYY.

In this way, Alembic's causality handles complexity and nonlinearity that traditional methods often overlook. So, while it doesn't give us the same kind of cause-and-effect that manipulation would, it's particularly effective in systems where we can't directly intervene and where non-linear interactions are at play.

Multi-channel marketing campaigns involving TV ads, social media, in-store displays, and other platforms introduce complex interdependencies that are difficult to untangle using traditional methods. For example, a TV ad may raise awareness, leading to social media engagement, which ultimately results in a purchase—traditional methods like A/B or random controlled trials testing struggle to model these interdependencies accurately.

Or, as we sometimes say to clients: "It's tough to A/B test a Super Bowl commercial, doing your largest B2B conference of the year, or naming a stadium."

## On Causal Chains - A Transfer from AI-Driven Drug Discovery

What distinguishes Alembic's approach is its ability to pinpoint common causal chains that have the greatest impact on revenue while filtering out less effective ones. Similar to techniques used in AI-driven drug discovery—where key molecular substructures are mapped to understand chemical properties[11]—Alembic applies this methodology to marketing.

The process of mining molecular substructures is aimed at finding frequent and significant patterns in chemical compounds. Similarly, Alembic evaluates the real-time interactions between various marketing touchpoints, continuously refining the causal graph. This allows marketers to not only assess the immediate impact of a marketing action but also see how each touchpoint fits within the broader multichannel ecosystem. This technique dynamically refines substructure searches in molecular data, Alembic adapts to evolving consumer behaviors, helping brands

[11] Borgelt, C., Meinl, T., & Berthold, M. (2005). MoSS: A Program for Molecular Substructure Mining. In Proceedings of the Fifth IEEE International Conference on Data Mining (pp. 123-130). University of Magdeburg.

allocate resources more effectively and ensuring that every marketing action contributes to revenue growth.

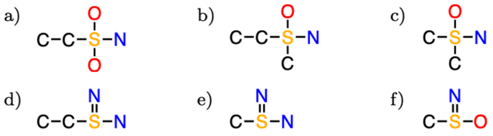


**Figure 1: A set of six example molecules.**

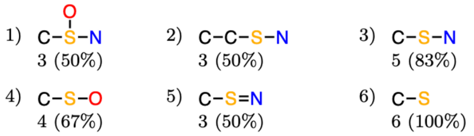


**Figure 3: The six frequent substructures that are found in the order in which they are generated.**

**Once the causal graph is constructed we can mine the best chain reactions and levers.**

**This is similar math to what is used discovering new compounds in medicine and perform molecular substructure mining.**

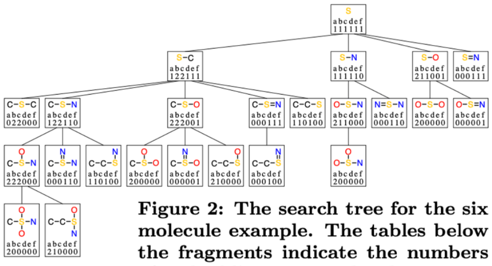
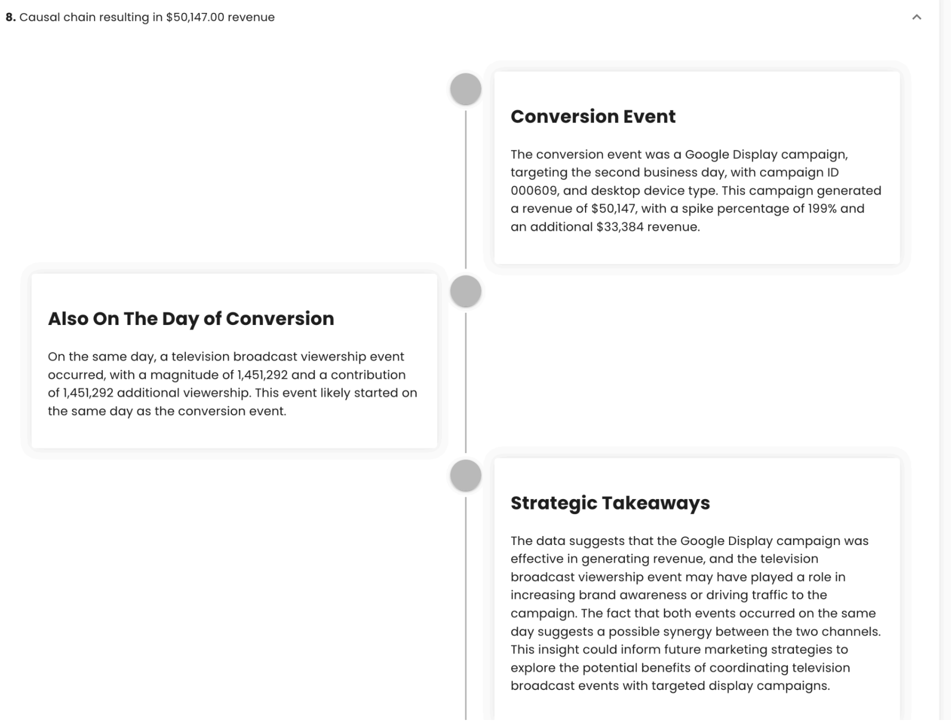


**Figure 2: The search tree for the six molecule example. The tables below the fragments indicate the numbers of embeddings per molecule.**

Much like molecular substructure mining in drug discovery, this method helps brands make smarter, more informed decisions by focusing on the patterns and connections that truly influence performance. By identifying and mapping these causal chains, Alembic enables marketers to focus on actions that consistently generate the highest returns.



**8.** Causal chain resulting in $50,147.00 revenue

### Conversion Event

The conversion event was a Google Display campaign, targeting the second business day, with campaign ID 000609, and desktop device type. This campaign generated a revenue of $50,147, with a spike percentage of 199% and an additional $33,384 revenue.

### Also On The Day of Conversion

On the same day, a television broadcast viewership event occurred, with a magnitude of 1,451,292 and a contribution of 1,451,292 additional viewership. This event likely started on the same day as the conversion event.

### Strategic Takeaways

The data suggests that the Google Display campaign was effective in generating revenue, and the television broadcast viewership event may have played a role in increasing brand awareness or driving traffic to the campaign. The fact that both events occurred on the same day suggests a possible synergy between the two channels. This insight could inform future marketing strategies to explore the potential benefits of coordinating television broadcast events with targeted display campaigns.

This method evaluates the immediate effects of a marketing action and allows marketers to see how each touchpoint fits into a broader multichannel ecosystem. By refining the graph in real-time, Alembic uncovers consumer behavior patterns that evolve as campaigns progress. This deeper understanding helps brands allocate resources more effectively across platforms, ensuring that every action contributes meaningfully to overall revenue growth.

By using causal chains that drive the most value and continuously discovering new opportunities, brands can make smarter decisions and achieve more targeted, revenue-driven outcomes.

### Why haven't people tried this before?

While advanced causal inference methods offer significant advantages, they have limitations that can be crippling to many business users.

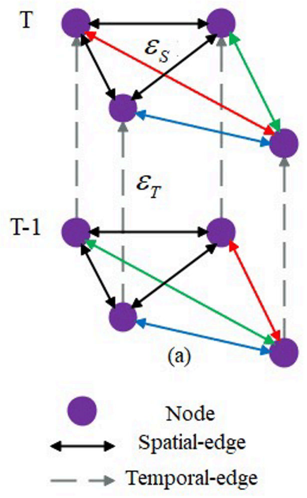
One major challenge is that these techniques require large datasets and immense computational power to function effectively. To meet this demand, Alembic owns one of the fastest supercomputers in the world, enabling us to handle such enormous data loads. For example, some implementations process over 100 billion rows of data annually. In general, inference is constrained not just by the problem itself but also by the amount of compute power you have available.[12]

Another hurdle is the availability of accurate and comprehensive data, particularly for cross-channel and cross-regional campaigns. Ensuring the quality and consistency of this data is critical for meaningful insights. To address this, Alembic has invested years in developing our sophisticated proprietary signal processor, ensuring we can seamlessly integrate and process complex datasets across diverse channels and regions.

## How do you store all this data in a format that works?

### Spatiotemporal Dynamic Graph

Understanding spatial and temporal relationships is critical in a multi-channel marketing environment. An evolution from DAG's[13] Alembic's Spatiotemporal Dynamic Graphs provide a detailed representation of customer interactions and marketing touchpoints, allowing marketers to model how these relationships change over time.



[12] Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science, 349(6245), 273–278

[13] Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. Retrieved from arXiv:1907.07271v2

Unlike traditional models that assume static, linear relationships between variables, spatiotemporal dynamic Graphs[14] enable a more nuanced analysis of the evolving connections between marketing actions and consumer behavior. This approach helps identify how different marketing activities influence outcomes across various channels and customer segments, offering a more accurate way to study the complexities of modern marketing ecosystems.

## Conclusion

Alembic's causal inference platform provides a powerful solution for navigating the complexities of modern marketing. As traditional methods like A/B testing and RCTs struggle to account for multi-channel interactions, Alembic uses spatiotemporal graph modeling to reveal the causal chains that drive consumer behavior. By identifying key pathways and evaluating how marketing efforts influence outcomes over time, Alembic helps marketers focus on actions that truly impact revenue while discovering new, effective strategies.

With the ability to process vast datasets, validate causal relationships, and adapt in real-time, Alembic offers actionable insights beyond surface-level analysis. This empowers marketers to make smarter, more proactive decisions by understanding how marketing actions fit into the broader multichannel ecosystem. In an increasingly fragmented consumer landscape, Alembic delivers the clarity needed to pinpoint the causal chains that matter most, transforming how brands interpret, predict, and optimize their marketing strategies.

[14] Zeghina, A., Leborgne, A., Le Ber, F., & Vacavant, A. (2024). Deep learning on spatiotemporal graphs: A systematic review, methodological landscape, and research opportunities. Neurocomputing, 594, 127861.

# Appendix: Common Questions About Alembic's Causal Inference for Data Scientists

This document addresses key questions about Alembic's causal inference platform, a new approach that challenges traditional marketing measurement techniques. As modern marketing grows more complex with multi-channel campaigns and non-linear interactions, conventional methods like A/B testing and attribution models often fall short. Alembic leverages advanced techniques, such as transfer entropy and temporal-spatial graph modeling, to offer a fresh perspective on uncovering causal relationships.

This Q&A tackles those concerns directly, covering topics like validating causal estimates, preventing overfitting, and handling large-scale datasets. We aim to provide clear answers and address the uncertainties around adopting causal inference in marketing, offering a grounded perspective on its potential benefits and challenges.

---

**Question 1:** *"Even if Alembic's causal can measure directional information flow, how do we ensure it's not picking up spurious relationships due to hidden common causes? Can it handle the presence of confounding variables?"*

**Answer:** This is a valid concern. Spurious relationships caused by hidden variables are a common challenge in any causal inference method. However, Alembic's causal is built with mechanisms to help mitigate this.

Alembic's causal is a time-directed measure, meaning it focuses on how information flows from the past of XXX to the future of YYY. This temporal aspect ensures that the direction of influence is clear—XXX cannot affect YYY retroactively, which helps avoid false correlations that are bidirectional or simultaneous.

Additionally, Alembic's causal operates conditionally, meaning it factors in YYY's past when measuring the influence of XXX. This is similar to how RCTs control for confounding variables—Alembic's causal controls for YYY's history to isolate the effect of XXX. However, like any observational method, it's not immune to hidden common causes. That's why it's often combined with other techniques, such as adding more variables to control for potential confounders, to strengthen its causal claims.

**Question 2:** *"Could you walk us through how you would estimate an efficiency curve for something like Google SEM spend?"*

**Answer:** Efficiency curves in isolation can be misleading, especially when externalities are involved. Marketing environments are interconnected, so looking at a single curve can overlook these dynamics.

For instance, a long-term sponsorship or external event (like a sports team making the World Series) can boost performance across multiple channels. Still, it wouldn't be captured in a standalone SEM efficiency curve. Instead, we focus on profitable revenue derived from the source and consider external factors like brand effects. Alembic doesn't replace traditional ROAS or efficiency metrics but complements them, offering more profound insights into externalities and profitable inventory.

**Question 3:** *"How do you validate the causal nature of the estimates your system produces?"*

**Answer:** Validating the causal nature of our system's estimates is crucial to our process with each client. Since Alembic is an enterprise-level solution, we prioritize collaboration with our clients during deployment, model training, and validation.

For client-specific validation, we use classic machine learning (ML) methods. This includes comparing predicted values to actual outcomes using confusion matrices and accuracy scores. We also implement back-testing, particularly during the tuning phase, where we integrate the required data streams.

Regarding the general approach, our system constructs a dynamic graph where nodes represent time-stamped events with attributes like magnitude or type, and edges represent both temporal and relational dependencies. The graph neural network (GNN) processes these relationships, learning from important temporal and relational patterns in the data. This allows the model to focus on the critical interactions that influence causality.

**Question 4:** *"I've read that Alembic's causal (Alembic's causal) requires large datasets to estimate information flow accurately. With Alembic's platform data—covering a wide range of marketing and sales touchpoints—how does data intensity impact the reliability of Alembic's causal in your system?"*

**Answer:** Great question. Alembic's causal (Alembic's causal) does indeed rely on large datasets to accurately capture the flow of information between variables. In smaller datasets, there's a risk of statistical noise affecting the results, potentially leading to less reliable insights.

However, Alembic is uniquely equipped to handle this challenge. Our platform ingests vast amounts of data across multiple touchpoints—ranging from high-level brand exposure (e.g., TV, podcasts, earned media) to more granular sales and CRM data. This diversity of data across time and channels ensures that we have the breadth and depth required for reliable Alembic's causal calculations.

Moreover, Alembic doesn't just rely on volume. The platform is built to continuously gather data in near real-time, allowing us to build robust datasets over time. This ensures that when we apply Alembic's causality, we're working with rich data that reflects actual marketing impacts, minimizing the risk of distortions caused by small sample sizes. Additionally, our preprocessing steps (e.g., time-series reconstruction and data cleaning) help filter out noise, ensuring that the input data is of high quality before Alembic's causality is applied.

In essence, the scale of data that Alembic processes ensures that Alembic's causal operates with a high level of reliability, leveraging the vast information flow from multiple marketing touchpoints.

---

**Question 5:** *"Alembic's causal is a powerful tool, but large models can run the risk of overfitting, especially in complex systems with many variables. Given Alembic's platform and its multi-channel, multi-variable environment, how do you mitigate the risk of overfitting when applying Alembic's causal?"*

**Answer:** You're absolutely right—like any powerful model, Alembic's causality could in theory overfit, particularly in complex environments where many variables interact simultaneously. Overfitting occurs when the model starts to capture the noise or quirks of the dataset rather than the actual underlying relationships, which can lead to misleading insights.

At Alembic, we mitigate the risk of overfitting in several ways:

1. **Cross-validation**: We use cross-validation techniques to test our models on different subsets of data. This ensures that the Alembic's causal model we build generalizes well to new, unseen data rather than overfitting to the specific dataset we're working with at any given time. It's critical to ensure robustness, especially when dealing with complex marketing environments where noise and outliers can distort the signal.
2. **Combining Alembic's causal with other methods**: While Alembic's causal is powerful for capturing non-linear, directional relationships, we often combine it with different statistical and causal inference methods to ensure that our results are consistent and not just the product of overfitting. By cross-referencing insights from multiple techniques (e.g., econometric models, regression analysis), we can validate the causal relationships we detect with Alembic's causal, ensuring they hold up under different analytical lenses.
3. **Dimensionality reduction and feature selection**: In complex systems with many variables, we apply dimensionality reduction techniques like principal component analysis (PCA) or factor analysis to identify the most critical features before applying

Alembic's causal. This helps prevent overfitting by focusing on the essential drivers of causality rather than trying to model every minor variable, which can introduce noise.

4. **Continuous monitoring and model updates**: One of Alembic's strengths is its ability to operate in near real-time. This allows us to continuously monitor the model's performance and adjust as new data comes in. If we detect that a model might be overfitting due to changing conditions or new patterns, we can quickly adjust the parameters or incorporate additional data to recalibrate the system.

---

**Question 6:** *"How should we think about the nature of the system's causal estimates? Are you estimating an average treatment effect or something related to heterogeneous treatment effects (HTE)?"*

**Answer:** That's an important distinction. Our system does account for heterogeneous treatment effects (HTE), which are crucial in complex environments like marketing, where different customer segments can respond to the same touchpoint in varied ways.

We model these complex interactions using spatial-temporal dynamic graphs, capturing how variables interact to produce different outcomes across subgroups. This ensures that relevant heterogeneity is adequately modeled and helps avoid issues such as multiple hypothesis testing errors.

In marketing, where sequential touchpoints matter, we focus on mapping how earlier interactions influence later ones, recognizing that responses can differ significantly across customer groups.

---

**Question 7:** *"Granger causality also looks at time-series data and measures directional influence. Why should we use Alembic's causal instead of Granger causality in complex systems?"*

**Answer:** Granger causality has been instrumental in establishing causality within linear systems, but the key difference is that it assumes a linear relationship between variables. This works well in many simple, well-behaved systems, but this assumption often breaks down when we move into more complex, non-linear environments—like financial markets or biological processes.

Alembic's causal doesn't make any assumptions about the nature of the relationship between XXX and YYY. It's fully non-parametric, meaning it can handle the kind of non-linear interactions we see in many real-world systems. This makes Alembic's causal particularly effective when linear models miss essential patterns. So, while Granger causality is excellent for linear dependencies, Alembic's causal provides a more flexible approach, especially in complex, dynamic environments.

**Question 8:** *"You mentioned some real-world applications of Alembic's causal. Can you provide specific examples where it's been successfully used to infer causality in practice?"*

**Answer:** Absolutely. Causal inferrence has been applied successfully in various fields where traditional methods struggle with non-linearity and complexity:

1. **Neuroscience**: This kind of causal has been used to measure the flow of information between different brain regions. For instance, researchers have applied Alembic's causal to understand how different brain parts communicate during cognitive tasks, identifying causal links that linear methods like correlation would miss.
2. **Climate Science**: In climate models, this type of causal has helped uncover how changes in sea surface temperature influence atmospheric conditions. These systems are highly interdependent and non-linear, making Alembic's causal ideal for tracing influence flow across different variables over time.
3. **Financial Markets**: This kind of causal has been used to measure how information from one stock market affects another, providing insights into cross-market influences. In these cases, the non-linear nature of market interactions makes Alembic's causal far more reliable than more straightforward methods.

**Question 9:** *"Are there any limitations to Alembic's causal? In large-scale data environments, could Alembic's causal suffer from overfitting or computational challenges?"*

**Answer:** Alembic's causal has some limitations, particularly regarding data and computational requirements.

1. **Data Intensity**: Alembic's causal requires large datasets to estimate the information flow between variables accurately. In smaller datasets, statistical noise can distort the results. In larger environments—like corporate data sets from marketing or financial systems—this isn't as much of a concern, but it's something to keep in mind.
2. **Computational Complexity**: Calculating Alembic's causal can be computationally intensive, especially over long periods or across many variables. However, advancements in computing

**Question 10:** *"How should we think about the nature of the system's causal estimates? Are you estimating an average treatment effect or something related to heterogeneous treatment effects (HTE)?"*

**Answer:** That's a great question and an important distinction. Alembic's system does account for heterogeneous treatment effects (HTE), which are especially relevant in marketing environments where different demographics and customer segments respond differently to the same touchpoints.

We model these complex interactions using spatial-temporal dynamic graphs incorporating demographic data and other variables. This enables us to capture how different subgroups experience varied outcomes, ensuring that the system properly reflects heterogeneity within the data. By modeling these differences, we reduce the risk of multiple hypothesis testing errors and ensure that our causal estimates are robust.

In marketing, sequential interactions matter significantly, and responses to touchpoints can vary widely based on factors like age or location. We focus on how earlier touchpoints influence later behaviors, ensuring that we can accurately map responses across different demographic groups.

---

**Question 11:** *“The system seems to rely heavily on temporal dependence to establish causality. How do you handle endogenous factors, like spending during holidays or demand-driven spending for brand campaigns?”*

**Answer:** That's a good point and one that we consider carefully. The system has a built-in general control for seasonality, allowing for identifying key holidays or events. However, we usually work closely with clients over the long term to incorporate the most relevant node features for their business needs. By doing this, the model learns the patterns related to endogenous spending and adjusts its predictions accordingly.

Simply reducing everything to temporal dependencies would overlook the complexity involved. We incorporate additional features to ensure the model can differentiate between seasonal effects and broader causal relationships.

---

**Question 12:** *“Once you run ETL and arrive at a 'geometric' dataset, what is the main advantage of using a temporal-spatial graph over other approaches?”*

**Answer:** The main advantage of *temporal-spatial graphs is* their ability to model both temporal and relational dependencies in a dynamic and complex environment. Traditionally, geometric data refers to physical distances between nodes, but in our case, we also consider time as a dimension.

*Temporal-spatial graphs* effectively capture the evolving relationships between events over time. This makes the model particularly powerful for predicting the future state of dynamic graphs compared to more static or linear approaches. By understanding both time and relational distance, the system is better equipped to analyze the true causal impact of marketing efforts.

**Question 1:** *“Why don’t you talk a lot about AI in this document.”*

**Question 13:** "Why don’t you talk much about AI in this document?"

**Answer:** "Alembic uses a range of AI, from LLMs to causal models, but we don’t feel the need to mention 'AI' every two sentences—especially when, at its core, all AI is just advanced computational statistics anyway."

# Methodology Testing and Validation

## Time-series Testing and Validation

For time-series we have a testing bench that we evaluate on standard pre-labeled datasets such as:

**Univariate:** 'AIOPS', 'Yahoo', 'WSD', 'NAB', 'TODS', 'UCR']

**Multivariate:** 'SMAP', 'MSL', 'SMD', 'SWaT'

While an anomaly detector should be evaluated in the with data from domain it is used in, we also benchmark it for general performance against other common time-series anomaly methods.

**Evaluation comparisons we run include:**

AR, LSTMADalpha, LSTMADbeta, AE, EncDecAD, SRCNN, Anomaly Transformer, Donut, FCVAE, SubLOF, SubOCSVM, Times Net and FITS.

## Causal Testing and Validation

We have tested with both real-world pre-labeled causal datasets and synthetic datasets.

One of the industry standards we use is Tigrimite[15]. This package is the industry and research standard developing and testing causal models[16].

Below are a selection of the tests we use to validate the Effective Renyi Transfer Entropy (ERTE) estimator. Each test targets a specific aspect of its performance and this includes handling high-dimensional data.  to detecting causal relationships and ensuring numerical stability. The table below provides a quick summary of these methods, their purpose, and implementation.

[15] Runge, J., Gerhardus, A., Varando, G. et al. Causal inference for time series. Nat Rev Earth Environ (2023). https://doi.org/10.1038/s43017-023-00431-y

[16] Generally: J. Runge (2018): Causal Network Reconstruction from Time Series: From Theoretical Assumptions to Practical Estimation. Chaos: An Interdisciplinary Journal of Nonlinear Science 28 (7): 075310. https://aip.scitation.org/doi/10.1063/1.5025050

# Causal Validity Tests Built in Tigrimite

| Method | What It Does | Why we should use | How It Works |
|---|---|---|---|
| Dimensionality Test | Checks ERTE behavior as data dimensionality increases. | Validates results reliability with increasing complexity | Tests finite TE values and expected degradation patterns |
| Small Sample Bias Test | Evaluates ERTE reliability with varying sample sizes. | Tries to understand how sample size affects reliability. | Measures variance and convergence across sample sizes. |
| Causal Chain Detection | Identifies direct and indirect causal relationships. | Validates causation and temporal ordering under noise. | Tests lag-specific TE values for known causal chains |
| Complex causality test | Detects feedback loops, multiple pathways/mixed lags | Verifies detection of complex causal relationships | Tests bi directional causation, direct/indirect effects, and relative strengths |
| Alpha Sensitivity Test | Assesses impact of Renyi alpha on dependency detection. | Validates estimator across linear and nonlinear dependencies. | Tests TE values across alpha values (e.g., 0.5 to 2.0) and verifies stability |
| Entropy Estimation Validation | Verifies entropy estimation across distributions | Verifies stability of entropy computation | Tests uniform, normal and bimodal distributions |
| Temporal Dependency Test | Verifies detection of lag-specific causal relationships | Ensures lag detection works | Tests for lag 2 dependencies and checks significance |
| Non-Stationarity Test | Analyzes TE under changing patterns in time series data. | Understand performance with non-stationary datasets. | Evaluates lag variations in time series |
| Numerical Stability Test | Validates TE against edge cases (e.g., sparse data). | Understand performance | Checks TE computations for sparse data + large ranges |

# Core Mathematical Validation Tests

## Dimensionality Test

**Context** - This test checks how well our ERTE estimation works when we increase the complexity/dimensions of the data.

### Why Is It Needed?

1. Reality Check: As we add more variables (dimensions), it becomes harder to find true relationships in the data.
2. Reliability Testing: We need to know when our results stop being trustworthy.

### Test Success Criteria

The test passes if:

1. All TE values are finite (no errors)
2. TE values decrease as dimensions increase
3. P-values increase as dimensions increase
4. Values stay within valid ranges (0-1 for p-values)

Note : The below function can be copied directly into our test_tceg_construction.py file as a new test method.

## Small Sample Bias Analysis

**Context** - Test that evaluates how sample size affects reliability of ERTE. TE in general can be sensitive to sample size

### Why Is It Needed?

1. Validates that the ERTE estimator works correctly with different sample sizes
2. Ensures estimates converge as sample size increases
3. Helps establish minimum sample size requirements for reliable results
4. Quantifies estimation uncertainty

### Test success criteria

- Variance decreases with increasing sample size
- Estimates converge to expected values

### What it tests

- Bias in small samples
- Convergence behavior
- Estimation variance
- Reliability thresholds

### Expected results

- Decreasing variance with sample size
- Convergence to true coupling value (0.5)
- P-values showing significance for larger samples
- Stable estimates above threshold (500 samples)

### Failure scenarios

- Variance increases with sample size
- No convergence to true value
- Excessive variation in reliable range

- True value outside confidence intervals

# Causal Structure Detection

## Linear Causal Chain

**Context -** This is a test for validating causal chain detection in the Transfer Entropy (TE) estimator. It creates a simple causal chain X -> Y -> Z where:

- X directly influences Y with lag 1
- Y directly influences Z with lag 2
- X indirectly influences Z through Y

**Why is it needed?**

- Validates the estimator can detect both direct and indirect causal relationships
- Ensures temporal ordering is preserved
- Tests noise tolerance
- Verifies that indirect effects are weaker than direct effects

**Test Success Criteria**:

- Significant TE values ($p < 0.05$) for direct connections
- Correct lag detection
- Indirect effects weaker than direct effects
- Preserved temporal ordering
- Noise tolerance

**What it Tests**:

- Direct causal detection
- Indirect causal detection
- Temporal ordering

**Expected Results**:

- X->Y: Strong causation, lag 1
- Y->Z: Moderate causation, lag 2
- X->Z: Weak causation, longer lag

**Failure Scenarios**:

- Missing direct connections
- Incorrect lag detection
- Indirect effect stronger than direct
- Loss of detection under noise
- Temporal ordering violations

## Complex Causal Structures

**Context -** A test for complex causal relationships in spike time series data. Tests the ERTE ability to detect various types of causal relationships

### Why is it needed?

- Validates ERTE estimator's ability to detect bidirectional (feedback) relationships
- Tests detection of multiple concurrent causal pathways (direct and indirect)
- Verifies accurate detection of different lag times in the same system
- Ensures proper handling of mixed temporal dependencies
- Validates relative strength estimation between different causal pathways

### Test success criteria

- Significant p-values (<0.05) for all known causal relationships (X↔Y, X→Z, Y→Z)
- Correct detection of specific lags (lag-1 X→Y, lag-2 Y→X, lag-2 X→Z, lag-1 Y→Z)
- Appropriate relative strengths (X→Y > Y→X, X→Z > Y→Z)
- Positive transfer entropy values for all causal relationships
- Detection of both direct and indirect pathways

### What it tests

- Bidirectional causation (X↔Y feedback loop)
- Direct causation (X→Z)
- Indirect causation (X→Y→Z)
- Multiple concurrent temporal dependencies
- Relative strength of different causal pathways
- System's ability to handle feedback loops
- Transfer entropy estimation with mixed temporal scales

### Expected results

- Significant TE from X->Y with lag 1
- Significant TE from X->Z with lag 2
- Higher TE values for direct connections vs indirect ones

### Failure scenarios

- False negatives in feedback loop detection (missing Y→X relationship)
- Incorrect lag identification in multi-pathway scenarios
- Failure to detect weaker indirect pathways (Y→Z)
- Numerical instabilities from feedback loop calculations
- Memory issues with long time series (T=1000)
- Sorting artifacts affecting temporal relationships
- Noise overwhelming weaker causal relationships
- False equivalence between direct and indirect pathway strengths
- Missing multiple relevant lags in feedback scenarios
- Temporal binning issues with mixed time scales

# Renyi-Specific Considerations

## Alpha Parameter Sensitivity

**Context -** A test for how different alpha values affect ERTE  estimation. Tests both linear and nonlinear dependencies

**Why is it needed?**

- Validates the robustness of TE estimation across different alpha values
- Ensures proper handling of different types of dependencies
- Helps detect numerical instabilities

**Test success criteria**

- All assertions pass
- Results are finite
- P-values are valid (0-1 range)
- Expected relationships between α and different dependency types hold

**What it tests**

- Linear dependencies (direct relationships)
- Nonlinear dependencies (squared relationships)
- Different alpha values (0.5 to 2.0)
- Numerical stability
- Statistical significance

**Expected results**

- Linear dependencies: significant TE near Nonlinear dependencies: better captured by alpha less than 1
- All values should be finite
- P-values should be meaningful

**Failure scenarios**

- Non-finite ERTE values
- Invalid p-values (outside 0-1)
- Unexpected relationships between alpha and dependency types
- Numerical instabilities at extreme alpha values

## Entropy Estimation Validation

**Context -** A test that validates the entropy estimation component by testing: Different probability distributions (uniform, normal, bimodal), Various Dirichlet prior values, Numerical stability, Expected entropy relationships

**Why is it needed?** The test ensures:

- Entropy estimation is reliable across different data distributions
- Dirichlet priors work as expected
- Results are numerically stable
- Basic entropy properties are maintained (e.g., uniform should have highest entropy)
- Edge cases don't cause numerical issues

**Test success criteria** - The test passes when:

- All ERTE values are finite numbers
- P-values are within [0,1]
- Entropy estimates remain stable across similar prior values (variation < 0.2)
- Results maintain expected relationships between distributions

**What it tests**

- Tests the following aspects of entropy estimation:
  Distribution types - Uniform (should have highest entropy) - Normal (should have lower entropy than uniform) - Bimodal (should have lower entropy than uniform)
  Prior values - Small priors (0.01) - Medium priors (0.1) - Large priors (0.5)
  Properties - Numerical stability - Prior sensitivity - Distribution relationships

**Expected results**

- All assertions pass
- ERTE values are finite
- P-values in [0,1]
- Small variation in entropy estimates for similar priors

**Failure scenarios**

- Non-finite ERTE values
- Invalid p-values
- High entropy variation with small prior changes
- Unexpected entropy relationships between distributions

# Time Series Specific Tests

## Temporal Dependency Structure

**Context -** A test that verifies the TransferEntropyEstimator can correctly detect temporal dependencies between time series

### Why is it needed?

- Validates that the TE estimator can identify causal relationships
- Ensures the lag detection mechanism works correctly
- Verifies the statistical significance testing is working

### Test success criteria

- Correctly identifies lag-2 dependency
- Returns statistically significant p-value (<0.05)
- Produces positive ERTE value

### What it tests

- Lag detection accuracy
- Statistical significance testing
- Basic temporal causality detection

### Expected results

- optimal_lags should include 2
- p_value should be < 0.05
- erte_value should be positive

### Failure scenarios

- Wrong lag detection (optimal_lags doesn't include 2)
- Non-significant result (p_value ≥ 0.05)
- Negative or zero ERTE value
- Numerical instabilities in time series
- Sorting issues with spike times

## Non-Stationarity Handling

**Context -** test for non-stationary time series analysis in transfer entropy estimation, specifically testing ‘regime’ changes

**Why is it needed?** To ensure the TE estimator can:

- Detect different coupling strengths
- Identify different temporal lags
- Handle transitions between regimes
- Correctly identify absence of coupling

**Test success criteria**

- Correct lag detection in different regimes
- Appropriate TE values for each regime
- Proper statistical significance
- Numerical stability

**What it tests**

- Strong coupling (lag 1)
- Weak coupling (lag 2)
- No coupling (random)
- Transitions between these regimes

**Expected results**

- Regime 1: High TE, significant p-value, lag 1
- Regime 2: Moderate TE, significant p-value, lag 2
- Regime 3: Near-zero TE, non-significant p-value

**Failure scenarios**

- Numerical instabilities in TE computation
- False positives in uncoupled regime
- Missed detection of true coupling
- Incorrect lag identification

# Edge Cases

## Numerical Stability

**Context -** Stability test for the Transfer Entropy estimator that checks how it handles different edge cases

### Why is it needed?

- To ensure the TE calculations remain stable under various challenging conditions
- To catch potential numerical issues before they affect production
- To validate the robustness of the algorithm

### Test success criteria

- All assertions pass
- No NaN or infinite values in results
- P-values stay within [0,1]
- Optimal lags are positive
- No runtime errors

### What it tests

- Sparse data handling (rare events)
- Large value ranges
- Precision boundary cases (very small numbers)

### Expected results

- All values should be finite
- P-values between 0 and 1
- ERTE values typically small but non-zero
- Optimal lags should be positive integers

### Failure scenarios

- NaN or infinite values in results
- P-values outside [0,1]
- Runtime errors from numerical instability
- Memory errors with large arrays
- Invalid optimal lag values

## Other Notes on Validation

**Cross-validation**: We use cross-validation techniques to test our models on different subsets of data. This ensures that the Alembic's causal model we build generalizes well to new, unseen data rather than overfitting to the specific dataset we're working with at any given time. It's critical to ensure robustness, especially when dealing with complex marketing environments where noise and outliers can distort the signal.

**Combining Alembic's causal with other methods**: While Alembic's causal is powerful for capturing non-linear, directional relationships, we often combine it with different statistical and causal inference methods to ensure that our results are consistent and not just the product of overfitting. By cross-referencing insights from multiple techniques (e.g., econometric models, regression analysis), we can validate the causal relationships we detect with Alembic's causal, ensuring they hold up under different analytical lenses.

# Bibliography

1. Del Mondo, G., Rodríguez, M. A., Claramunt, C., Bravo, L., & Thibaud, R. (2013). Modeling consistency of spatio-temporal graphs. *Data & Knowledge Engineering*, 84, 59-80.
2. Kohavi, R., Tang, D., Xu, Y., Hemkens, L. G., & Ioannidis, J. P. A. (2020). Online randomized controlled experiments at scale: lessons and extensions to medicine. *Trials*, 21(1), 150. https://doi.org/10.1186/s13063-020-4084-y
3. Velummailum, R. R., McKibbon, C., Brenner, D. R., Stringer, E. A., Ekstrom, L., & Dron, L. (2023). Data Challenges for Externally Controlled Trials: Viewpoint. *J Med Internet Res*, 25, e43484. https://doi.org/10.2196/43484
4. Hall, R. E. (1986). Market Structure and Macroeconomic Fluctuations. *Brookings Papers on Economic Activity*, 17(2), 285–338.
5. Benjamin-Chung, J., Arnold, B. F., Berger, D., Luby, S. P., Miguel, E., Colford, J. M., & Hubbard, A. E. (2018). Spillover effects in epidemiology: parameters, study designs and methodological considerations. *Int J Epidemiol*, 47(1), 332-347. https://doi.org/10.1093/ije/dyx201
6. Granholm, A., Alhazzani, W., Derde, L. P. G., et al. (2022). Randomised clinical trials in critical care: past, present and future. *Intensive Care Med*, 48, 164–178. https://doi.org/10.1007/s00134-021-06587-9
7. Velummailum, R. R., McKibbon, C., Brenner, D. R., Stringer, E. A., Ekstrom, L., & Dron, L. (2023). Data Challenges for Externally Controlled Trials: Viewpoint. *J Med Internet Res*, 25, e43484. https://doi.org/10.2196/43484
8. Hamaker, E. L., Mulder, J. D., & van IJzendoorn, M. H. (2020). Description, prediction and causation: Methodological challenges of studying child and adolescent development. *Developmental Cognitive Neuroscience*, 46, 100867. https://doi.org/10.1016/j.dcn.2020.100867
9. Scholtes, I. (2017). When is a Network a Network? Multi-Order Graphical Model Selection in Pathways and Temporal Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Nova Scotia, Canada. arXiv:1702.05499 [cs.SI].
10. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11), eaau4996. https://doi.org/10.1126/sciadv.aau4996
11. Borgelt, C., Meinl, T., & Berthold, M. (2005). MoSS: A Program for Molecular Substructure Mining. In *Proceedings of the Fifth IEEE International Conference on Data Mining* (pp. 123-130). University of Magdeburg.
12. Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245), 273–278. https://doi.org/10.1126/science.aac6076
13. Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. arXiv:1907.07271v2. https://doi.org/10.48550/arXiv.1907.07271
14. Zeghina, A., Leborgne, A., Le Ber, F., & Vacavant, A. (2024). Deep learning on spatiotemporal graphs: A systematic review, methodological landscape, and research opportunities. *Neurocomputing*, 594, 127861. https://doi.org/10.1016/j.neucom.2024.127861