RnR: a meta-solver for causal discovery in undersampled time series data

Mohammadsajad Abavisani*
Department of Electrical
and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332
s.abavisani@gatech.edu

Kseniya Solovyeva
Department of Computer Science
Georgia State University
Atlanta, GA 30303
ks.p.solo@gmail.com

David Danks
Department of Philosophy
University of Virginia
Charlottesville, VA 22903
eww3bh@virginia.edu

Vince Calhoun
Department of Electrical
and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332
vcalhoun@gatech.edu

Sergey Plis
Department of Computer Science
Georgia State University
Atlanta, GA 30303
s.m.plis@gmail.com

Abstract

Learning directed causal graphs from time-series data poses significant challenges, especially in fMRI where slow sampling rate obscures fast neural interactions. This temporal mismatch leads to undersampling, which can make multiple graphs equally plausible. We address this problem by explicitly modeling undersampling effects when recovering causal graphs. Our approach employs answer set programming (ASP) to enforce domain-specific constraints and optimize soft observational constraints, thereby identifying a Markov equivalence class for the resulting graph solutions. By customizing an ASP solver to collect multiple near-optimal solutions, we obtain not only the single best-fitting graph but an equivalence class of high-scoring graphs for expert consideration. This method, called Real-world noisy RASL (RnR), can also act as a meta-solver: it refines the output of other causal discovery algorithms by accounting for undersampling biases. In synthetic data and empirical brain network data, RnR produces more accurate causal graphs than state-of-the-art methods, improving F1 scores by an average of 46% over baseline methods when applied as a meta-solver, and by 64% on standalone synthetic data benchmarks. We demonstrate that RnR is robust to varying undersampling rates – maintaining high precision and recall even as sampling becomes more sparse – whereas baseline methods degrade significantly. Finally we test RnR on open-ended questions without knowing the ground truth like human brain fMRI data, showing that incorporating undersampling-aware constraints via ASP yields more

*https://sajadabvi.github.io

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

reliable and interpretable brain connectivity estimates than methods that ignore undersampling, bridging the gap between neural dynamics and observational data.[2]

1 Introduction

Causal discovery from functional Magnetic Resonance Imaging (fMRI) data has emerged as a critical endeavor to understand the neural mechanisms underlying cognitive processes and behaviors. Researchers not only seek to identify active brain regions during tasks, but also unravel the causal relationships between these regions, often referred to as "effective connectivity" [12]. Graphical causal models, such as causal Bayesian networks, have become a popular framework for this purpose, combining directed graphs with joint probability distributions to model the dependencies between different brain regions [33]. These models adhere to the Causal Markov Condition, which asserts that each node in a causal graph is conditionally independent of its non-descendants given its parents [43].

However, applying these models to fMRI data is fraught with challenges, particularly due to the mismatch between the temporal resolution of fMRI and the rapid timescale of neural processes. Typical sampling intervals in fMRI (one to three seconds) are much slower than the millisecond-level interactions between neurons, leading to significant undersampling [45]. This undersampling often results in multiple causal graphs being statistically indistinguishable given observed data, forming what is known as a Markov Equivalence Class [33, 43]. These problems are exacerbated by the indirect nature of the *Blood-Oxygen-Level Dependent* (BOLD)[3] signal, which reflects neural activity through complex and variable hemodynamic responses [18].

In addition, the inherent variability in the hemodynamic response across different brain regions and subjects, adds another layer of complexity. Variations in the time-to-peak of the BOLD response can lead to incorrect inferences about the direction of causality, particularly when using methods like Granger Causality (GC), which assumes a fixed temporal relationship between cause and effect [9]. Although some studies suggest that GC may be robust to certain variations in the hemodynamic response [39], the combination of measurement noise, undersampling, and hemodynamic variability often undermines the reliability of causal graphs learned from fMRI data.

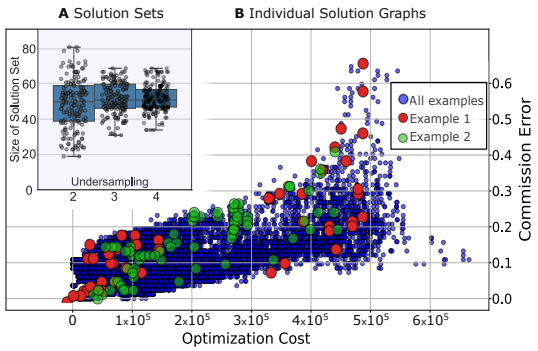

Figure 1: Top left(A): Size of optimization solution set across different undersamplings, repeated 100 times. Bottom right(B): Commission error of the solution vs. optimization cost for that solution. Solutions in one equivalent class are highlighted in red.

In response to these challenges, this paper proposes a novel approach called **RnR** (Real-world noisy RASL) that explicitly accounts for the effects of undersampling in the derivation of causal graphs. By employing constraint optimization through Answer Set Programming (ASP), we aim to identify the most probable causal graph from a set of potential candidates. ASP allows for the incorporation of domain-specific knowledge and constraints, facilitating the identification of not only a single graph but an equivalence class of possible graphs, thereby offering a more comprehensive understanding of the underlying causal structure [14]. We validate our approach using both simulated data and real fMRI data, demonstrating its superiority over existing methods. Our results suggest that ASP offers a powerful new tool for causal discovery in neuroimaging, providing more accurate and intuitive insights into the brain's functional connectivity.

[2]Code: https://github.com/neuroneural/gunfolds

[3]BOLD signal is the measurement obtained from fMRI that reflects changes in blood oxygenation related to neural activity. When neurons fire, blood flow to that region changes; fMRI captures these changes at a slow timescale (on the order of 1–2 seconds per sample). Importantly, the BOLD signal is an indirect and delayed proxy of neural firing.

2 Background

We study recovery of the true causal graph $\mathcal{G}^1$ from an observed, undersampled measurement graph $\mathcal{H}$, considering an unknown sampling rate u. We first fix notation and core facts, then state the limitation that motivates our approach.

A directed dynamic causal model extends standard causal models [33, 43] by adding time: n random variables $\mathbf{V} = \{V_1, \ldots, V_n\}$ appear at both the current timestep t ($\mathbf{V}^t$) and previous timesteps ($\mathbf{V}^{t-k}$). The full dynamic graph $\mathbf{G}$ is defined over $2\mathbf{V}$, and the only permissible edges are $V_i^{t-1} \to V_j^t$ (including $i = j$). We assume first-order Markov structure: $\mathbf{V}^t \perp\!\!\!\perp \mathbf{V}^{t-k} \mid \mathbf{V}^{t-1}$ for all $k > 1$.[4]

The underlying causal process evolves at timepoints $\{t^0, t^1, t^2, \ldots\}$. Data are *undersampled at rate* u when measurements occur only at $\{t^0, t^u, t^{2u}, \ldots\}$; in fMRI, slow sampling relative to neural dynamics makes $u > 1$ the norm [15]. We write $\mathbf{G}^1$ for the true causal graph and $\mathbf{G}^u$ for the same graph undersampled at rate u. To obtain $\mathbf{G}^u$, one "unrolls" $\mathbf{G}^1$ over u timesteps and marginalizes out all intermediate (unmeasured) variables.

Various representations have been developed for graphs with latent confounders, including maximal ancestral graphs (MAGs) [35] and partially-oriented ancestral graphs (PAGs) [46]. However, the latent structure induced by undersampling—unobserved intermediate states between sampled time points—is not equivalent to arbitrary latent confounding and is more naturally modeled using compressed (temporal) graph representations [8]. A compressed graph $\mathcal{G}$ represents a dynamic graph $\mathbf{G}$ over the n variables alone, with temporal information encoded implicitly in the edges: $V_i \to V_j$ in $\mathcal{G}$ iff $V_i^{t-1} \to V_j^t$ in $\mathbf{G}$. Undersampling is then a straightforward operation on compressed graphs: (1) $V_i \to V_j$ in $\mathcal{G}^u$ iff there is a directed path of length u from V_i to V_j in $\mathcal{G}^1$; and (2) $V_i \leftrightarrow V_j$ in $\mathcal{G}^u$ iff there exists V_k and a path length $s < u$ such that $\mathcal{G}^1$ contains directed paths of length s from V_k to both V_i and V_j (i.e., V_k is an unobserved common cause fewer than u timesteps back).

Given this framework, the causal discovery problem is: given an observed graph $\mathcal{H}$ inferred from data $\mathbf{D}$ at unknown undersample rate, find the equivalence class $[\![\mathcal{H}]\!] = \{\, \mathcal{G}^1 : \exists\, u \ (\mathcal{G}^u = \mathcal{H}) \,\}$. Since there are 2^{n^2} possible $\mathcal{G}^1$, this problem is NP-complete [8].

To address the structural implications of undersampling, several approaches have been developed. For instance, Danks and Plis [8] explored the problem structurally, while Gong et al. [15] provided a parametric approach for small systems. This line of work has been extended to more general subsampled and temporally aggregated VAR processes, including constraint-optimization methods for subsampled time series data [21], causal discovery from temporally aggregated time series [16], and recent proxy-variable approaches that can in principle identify the full causal graph from subsampled time series without parametric assumptions [27]. Other methods include LiNGAM [40], EC-GAN [26], amortized transformers [47], and continuous optimization [32]. None of these methods models undersampling: DYNOTEARS, for example, assumes the sampling rate is at least as fast as the causal process [32], and each returns a single graph on the measurement timescale. We therefore treat them as candidate inputs to RnR rather than as competitors; as a meta-solver, RnR can refine the graph any of them produces, as we show for MVGC, MVAR, GIMME and FASK in Figure 4. However, in recent years, the notion that supervised learning on synthetic data can bypass the fundamental theoretical restrictions of causal discovery has been challenged [30, 28, 24].

The Rate-Agnostic Structure Learning (RASL) algorithm [34] addressed the challenge of causal discovery from undersampled data by directly tackling undersampling while adopting a rate-agnostic approach that avoids assumptions about the undersampling rate. RASL was a big leap in computational efficiency. Its main drawback, however, was that it didn't scale well to larger graphs, leaving room for improvement. Building on these foundations, recent advancements in causal structure learning have led to the development of rate-agnostic approaches that significantly enhance the scalability and efficiency of these methods. The reformulation of RASL into a constraint satisfaction framework, expressed using a declarative language, has shown considerable promise by enabling more efficient analysis of large graphs called Solver-based RASL (sRASL) [1]. By incorporating constraints from strongly connected components (SCCs)[5] this approach offers a scalable and accurate method for causal discovery from undersampled data. It can analyze graphs with over 100 nodes—a

[4] This assumption is relatively weak, as we do not assume measurement at the causal timescale; the causal timescale could be arbitrarily fast. This is a form of causal sufficiency [43]. We also assume causal stationarity.

[5] Subsets of nodes in a directed graph where each node is reachable from every other node in the subset.

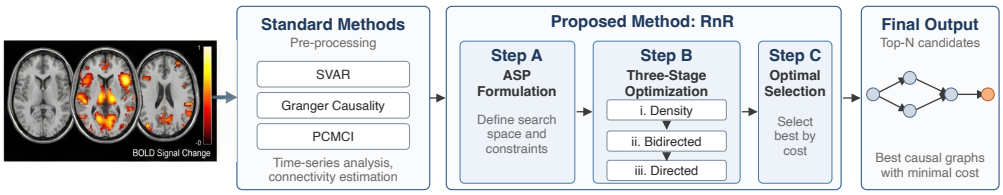

Figure 2: Overview of our pipeline. Starting from 4D fMRI volumes, classical methods extract a single "measured" graph from BOLD time series (e.g., using GC, SVAR, or PCMCI). However, these approaches typically ignore the effects of temporal undersampling. Our method extends beyond this step by accounting for the ambiguity introduced by slow sampling. Instead of selecting a single best-fit graph, we recover an equivalence class of plausible causal graphs that are consistent with both the observed data and the structural distortions caused by undersampling. Finally we present top N graphs based on optimization cost.

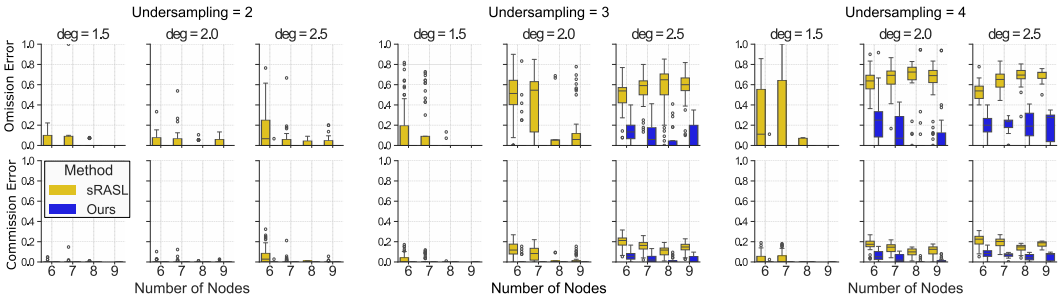

Figure 3: Normalized omission (top) and commission (bottom) errors for edge-breaking experiments with varying undersampling rates and graph densities, comparing the original approach with our improved sRASL-based method. Commission errors are false positives (spurious edges added); omission errors are false negatives (true edges missed).

significant improvement on earlier methods that struggled with much smaller graphs. The sRASL method improves solving time 1000-fold while maintaining the same theoretical guarantees as its predecessors. In parallel, Solovyeva et al. [42] showed that combining measurement graphs at multiple, deliberately undersampled rates can further shrink the equivalence class of system-timescale graphs and introduced an ASP-based solver (dRASL) that generalizes the RASL/sRASL encoding to multiple measurement timescales. However, sRASL shows limitations on real-world datasets, as the performance gains seen in simulations do not fully translate to noisy, practical conditions. One seminal work on the subject of causality and fMRI is [38], which demonstrated promising results on synthetic BOLD data. This study evaluates the robustness of such algorithms under constraints of undersampling.

Outside fMRI, related challenges arise when time series are irregularly sampled or acquired below the Nyquist rate. Bahadori and Liu [3] and Heerah et al. [20] adapt GC analysis to irregularly sampled biological time series, providing tests that explicitly account for unequal sampling intervals. In genomics, CaSPIAN and related “causal compressive sensing” approaches combine compressed sensing with GC to infer directed gene–gene interactions from heavily undersampled expression data [10]. Kathpalia and Nagaraj [22] further show that, under suitable structured sensing matrices, causal relations can be recovered directly from compressively sensed sparse signals. RnR fits into this broader methodological landscape by focusing on undersampling and hemodynamic distortion in fMRI, but using an ASP-based encoding that is, in principle, applicable to other subsampled or irregularly sampled modalities.

We benchmark a diverse set of widely used methods—including GIMME [13], MVGC [4, 5], MVAR [19], PCMCI [37], and FASK [38]—to capture a broad range of techniques, multivariate dependencies, and non-Gaussian orientation cues. By evaluating these baselines against our ASP-based approach (RnR) on undersampled data, we isolate the effects of sparse sampling and demonstrate that while FASK performs well among traditional baselines, RnR further enhances accuracy by making undersampling explicit; comparative results are further elaborated in Section 4.

3 Methods

In this section, we present Real-world noisy RASL (RnR),[6] our enhanced structure learning framework that builds upon the sRASL framework through five key innovations tailored for noisy, undersampled fMRI data. While each individual piece might appear relatively simple, they collectively have a large positive impact on causal structure learning from real-world data. The strength is in how these pieces fit together, each one shaped for a difficult, pressing scientific problem. More specifically, the RnR method: (1) uses a custom-made ASP solver to retrieve all plausible, near-optimal graphs within a cost threshold, rather than misleadingly offering only a single solution; (2) constrains the optimization with realistic edge density limits to ensure the resulting graphs are biologically plausible; (3) systematically addresses undersampling artifacts, thereby serving as a meta-solver that can refine the outputs of other causal discovery algorithms; (4) employs a prioritized, multi-stage process that sequentially matches graph density, resolves links, and orients edges in a way not previously used in causal structure learning; and (5) applies an adaptive weighting scheme to the cost function to discourage spurious connections while retaining high-confidence ones. Causal discovery from fMRI data has proven to be a notoriously difficult problem; RnR shows that real progress can be made through the intelligent integration of multiple innovations, all built on a theoretically solid foundation.

3.1 Overview of the sRASL Framework

The original sRASL framework [1] is an ASP-based approach that takes an initial directed graph $\mathcal{H}$ and applies structural constraints to refine it. We follow [1] and use Clingo[7] for our implementation. sRASL allows cycles within strongly connected components (SCCs), but requires the SCCs themselves to form a Directed Acyclic Graph (DAG). This DAG-over-SCCs structure prevents latent-level causal cycles. The original sRASL only returned a single optimal solution minimizing a cost function, but this posed limitations. Due to inherent uncertainty in the case of undersampling, the solution with the least cost is not necessarily the correct answer. Therefore, ignoring near-optimal alternatives in noisy data can be misleading.

3.2 Density Constraint for Realistic Graphs

Neuroscience knowledge can inform the expected density of functional brain networks. Empirical studies of brain connectivity (both functional and structural) show that brain graphs tend to be neither nearly empty nor fully connected; they often have a moderate density (for example, on the order of 10–30% of all possible edges, depending on node definition [41, 31, 2]. After running sRASL, we observed that some solutions were unrealistically sparse or dense, likely because the optimization only cared about matching $\mathcal{H}$ and not an absolute scale of connectivity. To correct this, we introduce an explicit density constraint in our ASP formulation. We conducted an ablation study to show the solutions are not overtly sensitive to the choice of this density prior (refer to Appendix C).

3.3 Optimization for Multiple Solutions Prioritized by Connection Type

To address the limitation of single-solution output, we modified Clingo to return all graphs G satisfying $C(G) \leq \Delta$, where $C(G)$ is the cost of graph G and Δ is a tolerance percentage relative to graph size. This ensures the method scales naturally with graph size, as a fixed integer tolerance would be overly restrictive for large graphs and too loose for small ones. Through hyperparameter tuning, we set $\Delta = 1.9 \cdot c_{\min}$, where $c_{\min}$ is the lowest cost in the solution set.(Details in Appendix B).

To efficiently recover causal graphs under severe noise and undersampling, we adopt a prioritized multi-stage optimization strategy that decomposes the total cost into structurally meaningful components. Our aim is to minimize the overall cost $C(G)$. We decompose the total cost as:

$$C(G) = C_o(G) + C_b(G) + C_d(G), \tag{1}$$

corresponding respectively to penalties on density (C_o), bidirected structure (C_b), and edge orientation (C_d). We minimize these costs in a prioritised, stepwise fashion. Each row below pairs the

[6] "RASL" is Rate-Agnostic Structure Learning, an approach introduced by Plis et al. (2015) for causal discovery without knowing the true sampling rate. Structural RASL (sRASL) from Abavisani et al. (2023) is discussed in Section 3.1.

[7] Clingo is an open-source solver for Answer Set Programming (ASP) that finds solutions (called "answer sets") to a given set of logical rules and constraints. In our context, Clingo explores different possible graphs and identifies those that satisfy all our constraints while optimizing a cost function. For a brief introduction to Clingo and ASP, please refer to Appendix F.

optimisation formulation (left) with the corresponding `clingo` weak-constraint encoding (right) that realises it, with feasible sets $\mathcal{F}_1 = \{G \in \mathcal{G} \mid C_o(G) \leq \epsilon_1\}$ and $\mathcal{F}_2 = \{G \in \mathcal{F}_1 \mid C_b(G) \leq \epsilon_2\}$.

$$\min_{G \in \mathcal{G}} C_o(G) = \lambda_o \cdot \left|\text{density}(G) - \text{density}(G^*)\right| \tag{2}$$

```
:~ abs_diff(D). [D*L_o@2, D]
```

$$\min_{G \in \mathcal{F}_1} C_b(G) = \lambda_b \cdot \#\{\text{bidirected mismatch to } G^*\} \tag{3}$$

```
:~ bidirected(X,Y,L),
   no_hbidirected(X,Y,W).
        [W*L_b @ 1, X,Y]
:~ not bidirected(X,Y,L),
   hbidirected(X,Y,W).
        [W*L_b @ 1, X,Y]
```

$$\min_{G \in \mathcal{F}_2} C_d(G) = \lambda_d \cdot \#\{\text{directed mismatch to } G^*\} \tag{4}$$

```
:~ directed(X,Y,L),
   no_hdirected(X,Y,W).
        [W*L_d @ 0, X,Y]
:~ not directed(X,Y,L),
   hdirected(X,Y,W).
        [W*L_d @ 0, X,Y]
```

where G denotes a candidate causal graph, $\mathcal{G}$ the space of all possible graphs, G^* the ground-truth. ϵ_1 and ϵ_2 are tolerance parameters. Priority weights $\lambda_o \gg \lambda_b \gg \lambda_d$ reflect the order of importance. The measurement graph $\mathcal{H}$ is encoded as ground ASP facts over the predicates `hdirected/3` and `no_hdirected/3`, whose leading `h` denotes their derivation from $\mathcal{H}$.

The lexicographic ordering is motivated by an *information hierarchy* among the three cost components, which are structurally asymmetric. Density (C_o) is a scalar global property—a single number constraining overall sparsity, easily satisfied by many graphs. Bidirected structure (C_b) encodes *which pairs* of nodes are confounded, an $O(n^2)$ pairwise property that significantly narrows the feasible set without specifying directionality. Directed orientation (C_d) is the highest-information component, encoding the full causal structure including direction. Optimizing in the order density $\rightarrow$ bidirected $\rightarrow$ directed corresponds to a coarse-to-fine refinement: each stage applies the lowest-information constraint that can still meaningfully prune the search space, leaving the high-information directional decisions for last, when the candidate set is already small. This mirrors the skeleton $\rightarrow$ v-structure $\rightarrow$ orientation pipeline of classical constraint-based causal discovery (e.g., the PC algorithm). The ablation in Appendix D.2 (Figure 11) confirms this intuition empirically: the solution set shrinks from $\sim$169,000 to $\sim$1,750 to $\sim$23 across the three stages, a four-order-of-magnitude reduction that would not be achievable by jointly optimizing all components.

We also confirmed the optimal hierarchy through systematic hyperparameter search (see Appendix A). For a more detailed discussion of the optimization and algorithmic treatment of RnR, please refer to Appendix D.

3.4 RnR as a Meta-Solver for Undersampling Effects

A significant challenge in causal discovery from fMRI is handling unobserved common causes (latent confounders) that arise from undersampling: if two brain regions A and B are both driven by an unmeasured region or stimulus that was not captured due to slow sampling, standard algorithms either link A–B with directed edges in a loop or miss the connection altogether. Bidirected edges ($A \leftrightarrow B$) explicitly represent such hidden-common-cause dependence. Some time-series methods (Granger causality [17], VAR models [29], PCMCI [37]) can produce bidirected or contemporaneous links, but many that ignore undersampling (standard PC, basic GC) cannot, potentially misrepresenting the true connectivity.

We address this by using RnR as a meta-solver that refines any initial graph $\mathcal{H}$ produced by a first-order method, with an augmentation motivated directly by the compressed-graph formalism in Section 2. Recall that a bidirected edge $V_i \leftrightarrow V_j$ in $\mathcal{G}^u$ is induced when $\mathcal{G}^1$ contains directed paths of length $s < u$ from some unobserved V_k to both V_i and V_j. The same dependence pattern (strong near-instantaneous coupling with no clear lag-direction) is exactly what drives first-order methods that ignore undersampling to output a mutual cycle $V_i \rightarrow V_j$ and $V_j \rightarrow V_i$: lacking a bidirected primitive, they split the dependence across both directions.

Our augmentation is therefore the structural *inverse* of this projection. Given an observed mutual cycle in $\mathcal{H}$, the structurally valid pre-images under the compressed-graph operator are: (i) a true bidirected edge representing latent confounding, (ii) a single directed edge in either direction, or (iii)

a genuine 2-cycle. For each mutual directed cycle in $\mathcal{H}$ we add a bidirected edge $A \leftrightarrow B$ alongside the two directed edges, with a low penalty weight w on the directed pair, reflecting the prior that mutual cycles under undersampling are more often confounding artifacts than genuine 2-cycles. The ASP solver then selects whichever combination—bidirected only, one directed edge, both, or all three—minimises the total cost given the remaining constraints, effectively injecting undersampling awareness into the upstream algorithm's output. When $\mathcal{H}$ already contains bidirected edges (e.g., from PCMCI), the augmentation is skipped. We demonstrate this capability in Section 4 by applying RnR to outputs from methods in [38], yielding improved results.

3.5 Implementation and Practical Considerations

Computational Efficiency: Constraint optimization with `Clingo` on our problem sizes (graphs of up to $N = 20$ nodes; Table 1) remained tractable, but because the problem is NP-complete, cost grows exponentially with N, and whole-brain parcellations with 50 or more regions are currently out of reach. However, our modular constraints (SCC-based DAG) significantly reduce complexity by disallowing many cyclic possibilities upfront. Further optimization or heuristic solving is a direction for future work, RnR's practical scope is therefore ROI-level analyses of up to about 20 regions, typical of effective-connectivity studies in fMRI.

Validation of Multiple Solutions: Importantly, retrieving multiple solutions allowed us to examine the stability of the recovered graph. We found that often many of the top solutions shared most edges in common, differing only in a few uncertain connections. This gave us confidence that our method identifies a core reliable structure plus a small set of alternative edges. We illustrate this in Figure 1A, which plots the number of solutions vs. undersampling rate: even at high undersampling, the solution set remains of modest size (tens of graphs), and in Figure 1B we show the relationship between solution cost and edge errors. Notably, we observed a positive correlation between optimization cost and edge commission error (false positive rate) across solutions. However, some higher-cost solutions had low commission error, indicating that near-optimal graphs can sometimes be almost as accurate as the optimal one. This justifies examining a small pool of top solutions. In practice, we found that the best solution (by ground-truth error) was often among the top 10 by cost. Therefore, our strategy is to take the top-N solutions from Clingo (we used $N = 10$) and then select the final graph by an additional criterion (e.g., best fit on a validation dataset or expert judgment). This approach yielded robust identification of the true causal graph in our tests.

4 Results

We replicated and extended the edge-breaking experiment from Abavisani et al. [1] to demonstrate the robustness of our approach on graphs with broken edges. In this experiment, we generated causal ground truth graphs $\mathcal{G}^1$ and undersampled them at various rates, then simulated noise by randomly deleting an edge. The goal was to test the ability of our method to recover the true graph structure, compared to the original sRASL. The results, depicted in Figure 3, show that our method consistently achieved lower omission and commission errors[8] compared to the original approach, even under high undersampling conditions. This illustrates that our improved algorithm can more effectively recover the true graph structure, including robustness against edge-breaking perturbations.

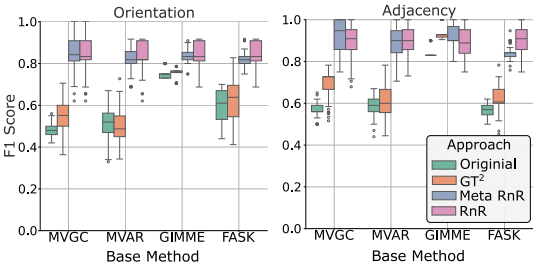

Figure 4: Performance comparison using Sanchez-Romero's data with and without the RASL meta-solver. Applying RnR on top of traditional methods improves accuracy by accounting for undersampling effects(purple), with additional gains achieved by using PCMCI+RnR (pink).

We further evaluated RnR on synthetic data from Sanchez-Romero et al. [38]. These data have been widely used within the neuroscience community, but were generated with small, simple graphs lacking real loops and using an idiosyncratic data generation method. We demonstrate that the methods that they study, including ones developed for these data (e.g., [38]), do not account for the effects of undersampling, which can bias causal discovery.

[8] A commission error is a false positive (i.e., incorrectly added edge). An omission error is a false negative (i.e., incorrectly omitted edge). We report these separately to characterize different types of errors (adding spurious links vs. missing real links).

Our approach applied RnR as a meta-solver: after running Sanchez-Romero's original methods, we applied RnR to the resulting graph to explicitly incorporate the effects of undersampling. This adjustment led to a 46% improvement in F1 on average, as RnR optimized the causal graph structure to better reflect the underlying dynamics. In essence, RnR served as an enhancement layer, correcting for undersampling effects ignored by previous methods (Figure 4).

Additionally, we explored the use of PCMCI [37] as an alternative to standard methods like SVAR and GC. Cook et al. [7] had previously demonstrated that SVAR and GC perform well for scenarios with isochronal bidirected edges, which arise due to undersampling, as well as directed edges. However, we observed that PCMCI performs better in these scenarios. Therefore, we incorporated PCMCI with RASL, achieving more accurate results by effectively handling undersampling and improving causal discovery. Figure 4 illustrates these improvements.

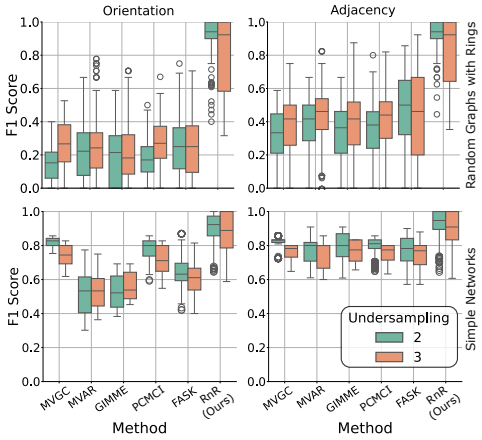

Figure 5: Comparison of Sanchez-Romero's simple data generation approach with larger, more complex VAR-generated graphs. Sanchez-Romero's data lacks real loops and is limited in scope, which may affect the generalizability of causal discovery results.

The inference problem solved by `RnR` is NP-complete in general, so any exact solver must scale exponentially in the number of nodes N. Table 1 gives an empirical sense of how that cost grows in practice. We follow the protocol of Abavisani et al. [1]: for each N we draw 10 random multi-SCC ring graphs (with maximum SCC size 6), simulate a stable vector-autoregressive process on each, convolve with the balloon haemodynamic model, and undersample to obtain BOLD-like time series. PCMCI is run on the resulting series and the measurement-scale graph it returns is fed to `RnR`. The table reports the median wall-clock time to find and prove the optimal causal-scale solution; the observed growth is consistent with the exponential bound, yet the method remains tractable at larger graph sizes representative of practical applications. Scaling further to the larger graphs required for full-brain applications remains future work. Hyttinen et al. [21] introduced the constraint-optimization formulation for noisy measurement graphs, demonstrated up to 12 nodes with u given; their satisfaction variant reaches 65 nodes, and sRASL [1] about 100 nodes on exact input. `RnR` keeps the edge-mismatch objective of Hyttinen et al. [21] and adds priority levels, a density term and meta-solver input.

Despite its widespread acceptance, Sanchez-Romero's data generation approach is limited in scope and complexity. It employs simple, small graphs without real loops and is generated using a specific, contrived process. To demonstrate the limitations of this dataset for broader causal discovery applications, we conducted experiments with larger VAR models on graphs with many variables and more complex structures, showcasing the limitations of Sanchez-Romero's setup (Figure 5). This underscores the importance of considering more realistic and complex data when testing causal discovery methods.

We also examined the effect of undersampling on time series data generated from a VAR model and then processed through a balloon model to simulate the BOLD response in fMRI data [6]. Given the smooth nature of the BOLD signal, undersampling by one-second intervals often leads to minimal information loss, while undersampling by larger intervals (e.g., two or three seconds) introduces significant inaccuracies. Figure 6 illustrates how different undersampling rates impact the preservation of temporal information in BOLD data, underscoring that aggressive undersampling can obscure meaningful connectivity patterns. Across these experiments (Figures 5, 6), RnR improves F1 by 64% on average over baselines.

4.1 Comparative Analysis of Causal Discovery Methods on fMRI Data

To validate our causal discovery framework and demonstrate its ability to extract novel information from neural imaging data, we conducted a comparative analysis using Granger Causality Mapping (GCM) [36] and our RnR. We used resting-state fMRI from the FBIRN Phase III study [23], decom-

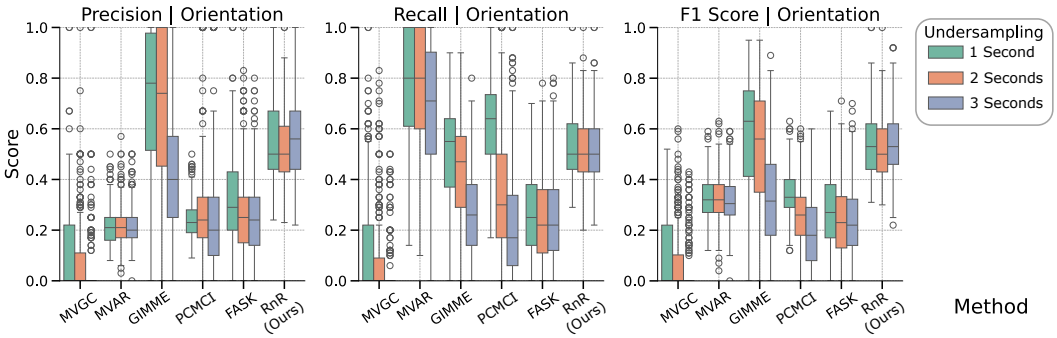

Figure 6: Impact of different undersampling rates (1s, 2s, 3s) on BOLD signal preservation in fMRI data simulated with the balloon model. Minimal error is observed with 1-second undersampling, whereas larger intervals degrade accuracy in all methods that don't account for the undersampling effect. Our method RnR accounts for this effect and does not suffer loss from undersampling.

posed with the NeuroMark_fMRI_1.0 ICA template [11], and selected the same six brain regions as [44], spanning the Central Executive Network (CEN), Salience Network (SN), and Default Mode Network (DMN). Data, preprocessing and estimator settings are given in Appendix G. By paying close attention to causal rates and sampling distinct causal properties, our results reveal that different methods can uncover complementary layers of network organization.

We first applied GCM to validate our pipeline against established findings. Our in-house implementation (Figure 7 left) showed substantial agreement with [44]'s results. Specifically, we confirmed the central role of the right fronto-insular cortex (rFIC) in mediating network interactions. As shown in Figure 7 left and Table 2, the strongest connection was the coupling between rFIC and the right dorsolateral prefrontal cortex (rDLPFC), replicating the key salience-executive link emphasized in the original study. Furthermore, we replicated the finding that the Salience Network exerts influence over the DMN, with robust directed connections from rFIC and ACC to DMN regions (VMPFC and PCC). These agreements validate the fidelity of our implementation and the quality of the underlying data processing.

N	median `RnR`
8	0.20 s
10	2.9 s
12	1.6 min
14	1.4 min
18	50 min
20	8.9 h

Table 1: `RnR` solve time on SCC-ring graphs (median over up to 10 instances per N).

Applying our RnR method to the exact same ICA components yielded a different but related set of answers, highlighting the method's sensitivity to structural rather than purely temporal causal signals. While GCM emphasized bidirectional coupling (rFIC $\leftrightarrow$ rDLPFC), RnR suggested a hierarchical structure driven by the Default Mode Network (Figure 7 right). The most prominent edge in RnR's output was a unidirectional connection from VMPFC to rFIC (Figure 7 right, Table 2), present in nearly 90% of solutions pooled across subjects. Because these data have no ground truth, we treat this edge as a hypothesis for follow-up rather than a validated discovery.

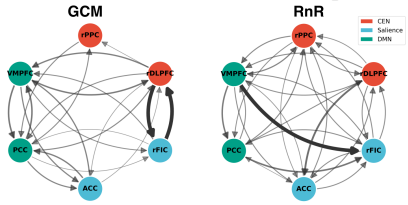

Figure 7: Comparative analysis of GCM and RnR. **Left:** GCM network showing strong bidirectional rFIC–rDLPFC coupling (thick black edges), replicating [44]'s salience–executive findings. **Right:** RnR network suggesting a dominant unidirectional pathway from the DMN, a candidate structural hierarchy absent in the bidirectional GCM results.

Comparing the two approaches reveals important convergences and one critical divergence. **Similarities:** Both methods agree on the high level of interconnectivity between the three networks and the central importance of the Salience Network (rFIC/ACC) in bridging distinct functional systems. **Key Difference:** The methods disagree on the primary driver of the system. GCM identifies the reciprocal temporal dynamics between rFIC and rDLPFC as the dominant feature, likely reflecting the active "switching" mechanism. In contrast, RnR suggests a structural precedence of VMPFC over rFIC, raising the hypothesis that the DMN's state may constrain or gate the activity of the Salience Network. Without ground truth, these data cannot settle which account is correct; the divergence is a testable hypothesis that accounting for undersampling brings into view.

Table 2: Top-5 candidate causal connections from GCM (temporal) and RnR (structural) on FBIRN data. Without ground truth, these are hypotheses, not validated connections.

GCM (temporal)			RnR (undersampling-aware)		
Connection	**Freq.**	**Interaction Type**	**Connection**	**Freq.**	**Interaction Type**
rFIC ↔ rDLPFC	58.1%	Salience ↔ CEN	VMPFC → rFIC	89.7%	DMN → Salience
VMPFC → PCC	39.4%	DMN (internal)	rDLPFC → ACC	67.3%	CEN → Salience
PCC → VMPFC	39.0%	DMN (internal)	PCC → rFIC	64.1%	DMN → Salience
VMPFC → rDLPFC	38.4%	DMN → CEN	VMPFC → PCC	62.6%	DMN (internal)
rDLPFC → VMPFC	37.4%	CEN → DMN	VMPFC → rPPC	59.1%	DMN → CEN

5 Conclusions and Future Directions

We addressed causal discovery from undersampled fMRI time series by encoding undersampling effects directly into an ASP-based optimisation, RnR. Rather than committing to a single point estimate, RnR returns an equivalence class of plausible causal graphs, surfacing the ambiguity that arises when slow, hemodynamically-distorted BOLD signals are used to recover fast neural dynamics. Empirically, RnR outperforms strong baselines on simulated data and yields meaningfully different connectivity patterns on real fMRI data, where it suggests a hierarchical, DMN-driven structure, a hypothesis for future validation, that is not visible to methods that ignore undersampling.

Limitations. RnR's output is only as good as its inputs. Recovery depends on the measurement graph $\mathcal{H}$ and on the density prior: a poor upstream estimator or a misspecified density range yields graphs that are internally consistent but not necessarily closer to the true structure (Appendix C). The size of the returned set depends on the tolerance Δ, which we set empirically (Appendix B). The meta-solver treats each mutual cycle in $\mathcal{H}$ as possible latent confounding and adds a bidirected candidate, a heuristic that can misattribute genuine reciprocal causation; this augmentation is skipped when $\mathcal{H}$ already contains bidirected edges. On real fMRI, where no ground truth exists, RnR's output should be read as hypotheses rather than validated connections. Finally, exact solving is exponential in N, so the current scope is ROI-level analyses of up to about 20 regions; extending it to whole-brain parcellations, and developing $\mathcal{H}$ estimators tailored to RnR, are natural next steps.

RnR aims to improve the reliability of causal discovery in neuroimaging, with potential benefits for basic neuroscience and, longer-term, biomarker discovery in psychiatric and neurodegenerative disorders; by exposing equivalence-class uncertainty rather than a single graph, it also encourages more honest uncertainty reporting in causal recovery. The principal risk is over-interpretation: edges recovered from undersampled BOLD data are constrained by inherent measurement limits, and individual connections should not be treated as guarantees of mechanistic causation. We release RnR as a research and hypothesis-generation tool, not a clinical decision-support system.

Acknowledgments and Disclosure of Funding

This work was supported by NIH R01MH129047, and in part by NSF 2112455 and NIH 2R01EB006841. FBIRN data collection was supported by the National Center for Research Resources at the National Institutes of Health (NIH 1 U24 RR021992 and NIH 1 U24 RR025736).

The authors declare no competing interests.

References

[1] Mohammadsajad Abavisani, David Danks, and Sergey Plis. Grace-c: Generalized rate agnostic causal estimation via constraints. In *The Eleventh International Conference on Learning Representations*, 2023.

[2] Georgette Argiris, Yaakov Stern, and Christian Habeck. Brain resilience to targeted attack of resting bold networks as a measure of cognitive reserve. *Imaging Neuroscience*, 2025.

[3] Mohammad Taha Bahadori and Yan Liu. Granger causality analysis in irregular time series. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 660–671. SIAM, 2012. doi: 10.1137/1.9781611972825.57.

[4] Lionel Barnett and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical Review Letters*, 103(23):238701, 2009.

[5] Lionel Barnett and Anil K Seth. The mvgc multivariate granger causality toolbox: a new approach to granger-causal inference. *Journal of neuroscience methods*, 223:50–68, 2014.

[6] Richard B Buxton, Eric C Wong, and Lawrence R Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic resonance in medicine*, 39(6):855–864, 1998.

[7] John W Cook, David Danks, and Sergey M Plis. Learning dynamic structure from undersampled data. In *Proceedings of the UAI Causality Workshop*, 2017.

[8] David Danks and Sergey Plis. Learning causal structure from undersampled time series. In *NIPS Workshop on Causality*, volume 1, pages 1–10, 2013.

[9] Olivier David, Isabelle Guillemain, Sandrine Saillet, Sebastien Reyt, Colin Deransart, Christoph Segebarth, and Antoine Depaulis. Identifying neural drivers with functional mri: an electrophysiological validation. *PLoS biology*, 6(12):e315, 2008.

[10] Mo Deng, Amin Emad, and Olgica Milenkovic. Causal compressive sensing for gene network inference. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 696–699. IEEE, 2012.

[11] Yuhui Du, Zening Fu, Jing Sui, Shuang Gao, Ying Xing, Dongdong Lin, Mustafa Salman, Anees Abrol, Md Abdur Rahaman, Jiayu Chen, L. Elliot Hong, Peter Kochunov, Elizabeth A. Osuch, and Vince D. Calhoun. Neuromark: An automated and adaptive ica based pipeline to identify reproducible fmri markers of brain disorders. *NeuroImage: Clinical*, 28:102375, 2020. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2020.102375. URL https://www.sciencedirect.com/science/article/pii/S2213158220302126.

[12] Karl J Friston. Functional and effective connectivity in neuroimaging: a synthesis. *Human brain mapping*, 2(1-2):56–78, 1994.

[13] Kathleen M. Gates and Peter C.M. Molenaar. Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, 63(1):310–319, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2012.06.026. URL https://www.sciencedirect.com/science/article/pii/S1053811912006404.

[14] Martin Gebser, Benjamin Kaufmann, and Torsten Schaub. Conflict-driven answer set solving: From theory to practice. *Artificial Intelligence*, 187:52–89, 2012.

[15] Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In *International Conference on Machine Learning*, pages 1898–1906. PMLR, 2015.

[16] Mingming Gong, Kun Zhang, Bernhard Schölkopf, Clark Glymour, and Dacheng Tao. Causal discovery from temporally aggregated time series. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, pages 1066–1075. AUAI Press, 2017.

[17] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[18] Daniel A Handwerker, John M Ollinger, and Mark D'Esposito. Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, 21(4):1639–1651, 2004.

[19] L Harrison, William D Penny, and Karl Friston. Multivariate autoregressive modeling of fmri time series. *Neuroimage*, 19(4):1477–1491, 2003.

[20] Sachin Heerah, Roberto Molinari, Stéphane Guerrier, and Amy Marshall-Colon. Granger-causal testing for irregularly sampled time series with application to nitrogen signalling in arabidopsis. *Bioinformatics*, 37(16):2450–2460, 2021. doi: 10.1093/bioinformatics/btab126.

[21] Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. A constraint optimization approach to causal discovery from subsampled time series data. *International Journal of Approximate Reasoning*, 90:208–225, 2017.

[22] Aditi Kathpalia and Nithin Nagaraj. Granger causality for compressively sensed sparse signals. *Physical Review E*, 107(3):034308, 2023. doi: 10.1103/PhysRevE.107.034308.

[23] David B. Keator, Theo G.M. van Erp, Jessica A. Turner, Gary H. Glover, Bryon A. Mueller, Thomas T. Liu, James T. Voyvodic, Jerod Rasmussen, Vince D. Calhoun, Hyo Jong Lee, Arthur W. Toga, Sarah McEwen, Judith M. Ford, Daniel H. Mathalon, Michele Diaz, Daniel S. O'Leary, H. Jeremy Bockholt, Syam Gadde, Adrian Preda, Cynthia G. Wible, Hal S. Stern, Aysenil Belger, Gregory McCarthy, Burak Ozyurt, and Steven G. Potkin. The function biomedical informatics research network data repository. *NeuroImage*, 124:1074–1079, 2016. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.09.003. URL https://www.sciencedirect.com/science/article/pii/S1053811915007995. Sharing the wealth: Brain Imaging Repositories in 2015.

[24] Hebi Li, Qi Xiao, and Jin Tian. Supervised whole dag causal discovery, 2020. URL https://arxiv.org/abs/2006.04697.

[25] Vladimir Lifschitz. Answer set programming and plan generation. *Artificial Intelligence*, 138 (1-2):39–54, 2002.

[26] Jinduo Liu, Junzhong Ji, Guangxu Xun, Liuyi Yao, Mengdi Huai, and Aidong Zhang. Ec-gan: Inferring brain effective connectivity via generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4852–4859, 2020.

[27] Mingzhou Liu, Xinwei Sun, Lingjing Hu, and Yizhou Wang. Causal discovery from subsampled time series with proxy variables. In *Advances in Neural Information Processing Systems*, 2023.

[28] Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. *Advances in Neural Information Processing Systems*, 35: 13104–13118, 2022.

[29] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[30] Francesco Montagna, Max Cairney-Leeming, Dhanya Sridhar, and Francesco Locatello. Demystifying amortized causal discovery with transformers. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=9Lgy7IGSfp.

[31] Sarah E Morgan, Sophie Achard, Maite Termenon, Edward T Bullmore, and Petra E Vértes. Low-dimensional morphospace of topological motifs in human fmri brain networks. *Network Neuroscience*, 2(02):285–302, 2018.

[32] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. Pmlr, 2020.

[33] Judea Pearl. *Causality*. Cambridge university press, 2009.

[34] Sergey Plis, David Danks, Cynthia Freeman, and Vince Calhoun. Rate-agnostic (causal) structure learning. *Advances in neural information processing systems*, 28, 2015.

[35] Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

[36] Alard Roebroeck, Elia Formisano, and Rainer Goebel. Mapping directed influence over the brain using granger causality and fmri. *Neuroimage*, 25(1):230–242, 2005.

[37] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11):eaau4996, 2019. doi: 10.1126/sciadv.aau4996. URL https://www.science.org/doi/abs/10.1126/sciadv.aau4996.

[38] Ruben Sanchez-Romero, Joseph D Ramsey, Kun Zhang, Madelyn RK Glymour, Biwei Huang, and Clark Glymour. Estimating feedforward and feedback effective connections from fMRI time series: Assessments of statistical methods. *Network Neuroscience*, 3(2):274–306, 2019.

[39] Anil K Seth, Paul Chorley, and Lionel C Barnett. Granger causality analysis of fmri bold signals is invariant to hemodynamic convolution but not downsampling. *Neuroimage*, 65:540–555, 2013.

[40] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Marko Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

[41] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

[42] Kseniya Solovyeva, David Danks, Mohammadsajad Abavisani, and Sergey Plis. Causal learning through deliberate undersampling. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 518–530. PMLR, 2023.

[43] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[44] Devarajan Sridharan, Daniel J Levitin, and Vinod Menon. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences*, 105(34):12569–12574, 2008.

[45] Pedro A Valdes-Sosa, Alard Roebroeck, Jean Daunizeau, and Karl Friston. Effective connectivity: influence, causality and biophysical modeling. *Neuroimage*, 58(2):339–361, 2011.

[46] Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9: 1437–1474, 2008.

[47] Zuozhen Zhang, Ziqi Zhang, Junzhong Ji, and Jinduo Liu. Amortization transformer for brain effective connectivity estimation from fmri data. *Brain Sciences*, 13(7), 2023. ISSN 2076-3425. doi: 10.3390/brainsci13070995. URL https://www.mdpi.com/2076-3425/13/7/995.

Appendix

A Hyperparameter Optimization for RnR Priorities

A.1 Motivation and Methodology

Our RnR algorithm employs a weighted optimization framework where five priority parameters control the relative importance of different constraints in the Clingo answer set programming solver. These priorities, denoted as $\mathbf{p} = [p_1, p_2, p_3, p_4, p_5]$, govern the trade-offs between: (1) minimizing false positive directed edges (p_1), (2) minimizing false positive bidirected edges (p_2), (3) minimizing false negative directed edges (p_3), (4) minimizing false negative bidirected edges (p_4), and (5) matching the expected graph density (p_5). Each priority can take integer values from 1 (lowest) to 5 (highest), resulting in $5^5 = 3{,}125$ possible configurations[9]. In Equation 1, $C_o(G)$ corresponds to p_5, $C_b(G)$ corresponds to p_2, p_4, and $C_d(G)$ corresponds to p_1, p_3. Crucially, the aggregate parameters λ_o, λ_b, and λ_d in Equations $2-4$, derived from these components do not function as continuous coefficients typical in gradient-based learning. Instead, they represent distinct optimization priority levels that enforce a strict lexicographic ordering within the ASP solver. Consequently, the systematic tuning of $\mathbf{p}$ establishes the hierarchical relationship between these terms—determining, for instance, whether satisfying biological density constraints (λ_o) takes precedence over resolving edge orientation (λ_d). Thus, this optimization process defines the solver's structural decision logic rather than adjusting scalar penalty magnitudes.

To systematically identify the optimal priority configuration for causal direction recovery, we conducted an exhaustive hyperparameter search using high-performance computing resources. The search was designed to maximize orientation F1 score, which measures the accuracy of recovered edge directions—the most critical metric for causal discovery. We employed a SLURM-based parallel computing framework where each of the 3,125 priority configurations was evaluated independently on a computing cluster. For each configuration, we generated data from a known ground truth causal graph (simple networks from [38]) and evaluated performance across five independent batches to ensure statistical robustness. The PCMCI algorithm with partial correlation tests was used for initial graph estimation, followed by RnR. Performance metrics were computed by comparing the recovered causal-timescale graph against the ground truth.

A.2 Experimental Results

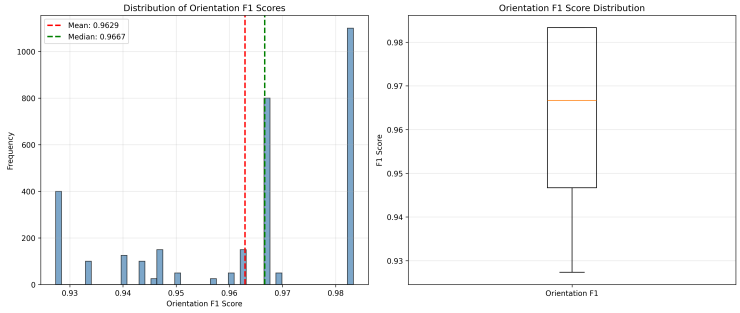

Figure 8: Distribution of orientation F1 scores across all 3,125 tested priority configurations. The distribution exhibits bimodal behavior with a primary mode near 0.98 and a secondary mode near 0.93. Red dashed line: mean (0.9629); green dashed line: median (0.9667).

The exhaustive search revealed substantial variation in performance across priority configurations (Figure 8). Orientation F1 scores ranged from 0.927 to 0.983, with a mean of 0.963 ± 0.020 (mean $\pm$ std) and median of 0.967. The distribution exhibited a bimodal pattern, with a dominant mode of high-performing configurations (F1 ≈ 0.98) and a secondary cluster of lower-performing configurations (F1 ≈ 0.93), suggesting distinct regions of the priority space correspond to qualitatively different optimization behaviors.

[9] In practice, possible configurations are much less, since for example $[1, 1, 1, 1, 2]$ and $[1, 1, 1, 1, 5]$ are practically the same. Our hyperparameter tuning experiments confirms this finding as well.

The optimal priority configuration was identified as $\mathbf{p}^* = [1, 2, 1, 2, 3]$, achieving an orientation F1 score of 0.983. This configuration prioritizes maintaining correct graph density ($p_5 = 3$) above all other constraints, while assigning moderate importance to matching bidirected edges ($p_2, p_4 = 2$) and lower importance to directed edge constraints ($p_1, p_3 = 1$).

The systematic hyperparameter optimization yielded a 2.1% improvement in orientation F1 score over the baseline configuration (from 0.963 to 0.983), corresponding to a 56% reduction in orientation errors. This improvement, while seemingly modest, is significant given the already high baseline performance and the critical nature of causal direction accuracy in downstream analyses that rely on the recovered graph.

A.3 Compute Resources

All experiments were run on an internal SLURM-managed CPU cluster (no GPUs were used; clingo's ASP solver is CPU-only). Each individual job was allocated **16 CPU cores** and a memory footprint scaled to the graph size: **8 GB** for $N \in \{8, 10\}$, **32 GB** for $N \in \{12, 14\}$, **64 GB** for $N \in \{18, 20\}$. Each job was given a 24-hour walltime; `RnR` was internally bounded by a 35-hour timeout to leave a 1-hour margin for log writing.

B Hyperparameter Optimization for Solution Selection Threshold

B.1 Motivation and Methodology

Our RnR algorithm typically returns multiple candidate solutions with associated optimization costs. Rather than selecting a single minimum-cost solution, averaging across multiple near-optimal solutions can improve robustness by mitigating the impact of optimization noise and local minima. The critical question becomes: which solutions should be included in this ensemble?

We introduce a cost-based selection threshold, δ, defined as a percentage of the minimum solution cost. Specifically, given a set of candidate solutions $\mathcal{S} = \{(G_i, c_i)\}$ where G_i represents a causal graph and c_i its optimization cost, we select the subset:

$$\mathcal{S}_\delta = \{(G_i, c_i) \in \mathcal{S} : c_i \leq c_{\text{min}} + \delta \cdot c_{\text{min}}\}$$

where $c_{\text{min}} = \min_i c_i$. The parameter δ thus controls the solution selection bandwidth: $\delta = 0$ selects only the minimum-cost solution, while larger δ values include progressively more solutions with higher costs. The final causal graph estimate is obtained by averaging orientation F1 scores across all solutions in $\mathcal{S}_\delta$.

To identify the optimal threshold, we conducted a systematic hyperparameter search testing 21 values of $\delta \in [0\%, 200\%]$ in 10% increments. Each threshold was evaluated on simple networks similar to previous section, with undersampling rate 3 across five independent batches for statistical robustness. For each batch, PCMCI with partial correlation tests generated an initial graph estimate. The orientation F1 score—measuring edge direction accuracy—was computed for all solutions within the δ threshold and averaged to obtain the final performance metric.

B.2 Experimental Results

The hyperparameter search revealed a clear optimal threshold at $\delta^* = 90\%$ of the minimum cost, achieving an orientation F1 score of 0.7996 (Figure 9, left). Performance exhibited moderate sensitivity to the threshold choice, with F1 scores ranging from 0.7940 to 0.7996 across all tested values—a variation of approximately 0.7%. The optimization landscape showed a distinct peak at $\delta^* = 90\%$, with performance degrading for both smaller thresholds ($\delta < 90\%$) and larger thresholds ($\delta > 100\%$).

Precision and recall metrics remained relatively stable across the threshold range (Figure 9, right), with precision hovering around 0.790 and recall around 0.802. This stability suggests that the threshold primarily affects which near-optimal solutions are averaged rather than fundamentally altering the precision-recall trade-off. The slight decline in precision at very high thresholds ($\delta > 150\%$) indicates that including solutions with costs substantially above the minimum begins to introduce lower-quality predictions.

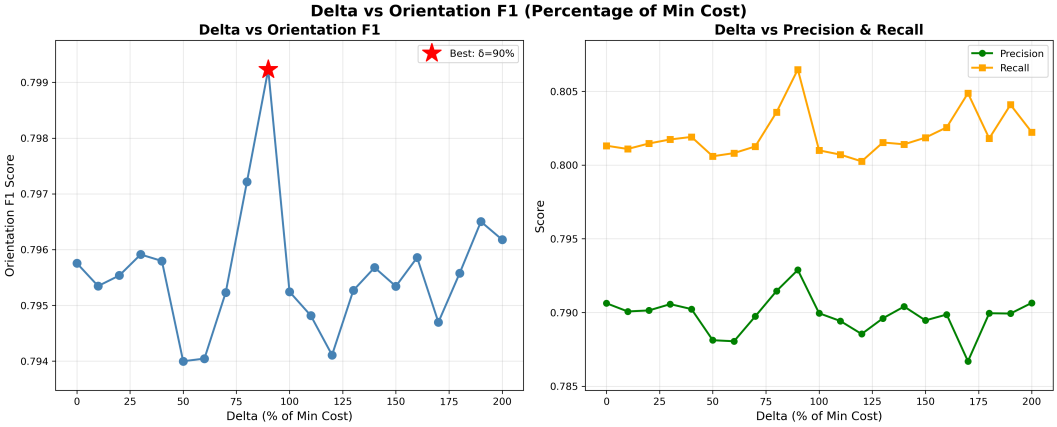

Figure 9: Hyperparameter optimization for solution selection threshold δ. **Left:** Orientation F1 score versus δ (percentage of minimum cost). Red star indicates optimal threshold at $\delta^* = 90\%$, achieving F1 = 0.7996. **Right:** Precision and recall trends show stability across thresholds with slight precision decrease at higher δ values.

B.3 Analysis and Implications

The optimal threshold $\delta^* = 90\%$ represents a balance between solution quality and ensemble diversity. At this threshold, solutions with costs up to 1.9 times the minimum are included. This relatively broad inclusion criterion suggests that RnR's cost function, while effective at guiding optimization, may not perfectly correlate with orientation accuracy—solutions with moderately higher costs can still contribute positively to the ensemble average.

The percentage-based formulation of δ provides generalizability across different networks and problem scales. Unlike an absolute cost threshold, which would require recalibration for each experimental configuration, the percentage-based approach adapts automatically to the characteristic cost scale of each problem. This scale-invariance is particularly valuable for automated analysis pipelines processing heterogeneous datasets.

Comparing solution selection strategies, the ensemble approach with $\delta^* = 90\%$ (F1 = 0.7996) offers a modest but consistent improvement over single-solution selection ($\delta = 0\%$, F1 = 0.7958), corresponding to a 1.9% reduction in orientation errors. This improvement comes at minimal computational cost, as all candidate solutions are already generated during RnR optimization.

The relatively flat performance plateau for $\delta \in [70\%, 110\%]$ (F1 variation $< 0.2\%$) indicates that the optimization is robust to moderate threshold misspecification. This tolerance provides practical flexibility in deployment scenarios where precise threshold tuning may be infeasible. All subsequent analyses employ the optimized solution selection threshold $\delta^* = 90\%$ of the minimum cost.

C Sensitivity Analysis of the Density Prior

Motivation. The cost component $C_o(G) = \lambda_o \cdot |\text{density}(G) - \text{density}(G^*)|$ in our prioritised optimisation requires a target density d^* for the (unknown) ground-truth graph G^*. In practice d^* is unavailable and must be supplied as a prior — e.g., from the measurement graph $\mathcal{H}$, from a population-level estimate, or from domain knowledge. A natural question is therefore how sensitive RnR is to misspecification of d. This appendix reports a controlled sweep designed to answer that question.

Experimental setup. For each ground-truth graph we evaluate RnR at five density priors,

$$d_{\text{assumed}} = \alpha \cdot d_{\text{true}}, \qquad \alpha \in \{0.7,\ 0.9,\ 1.0,\ 1.1,\ 1.3\},$$

corresponding to misspecification offsets of $\{-30\%, -10\%, 0\%, +10\%, +30\%\}$, with 0% being the oracle setting. Ground-truth graphs are produced by the `ringmore` construction over $n \in \{5, 6\}$ nodes with $\{1, 2, 3\}$ random extra edges. From each graph we simulate a stable VAR process with

5,000 samples and Gaussian noise ($\sigma = 0.1$), then undersample at rates $u \in \{2, 3\}$. PCMCI ($\tau_{\max}$=2, default α) is run on the undersampled series to obtain the measurement graph $\mathcal{H}$ that is fed into RnR. Each configuration (n, extra, u) is repeated over 10 random seeds. Every RnR call uses our canonical priority weights $[1, 2, 1, 2, 3]$, `optim=optN`, and a 2-hour wall-clock cap. For each $(n, \text{extra}, u, \text{seed})$ instance, the *same* simulated time series is reused across all five priors so that any change in F1 is attributable to the prior alone, not to data variability.

Results. Figure 10 reports the orientation and adjacency F1 of the lowest-cost RnR solution as a function of the misspecification offset, aggregated across all configurations and seeds. Two observations stand out:

- **Robustness.** Both F1 curves are nearly flat across the full $\pm 30\%$ range. The variation between settings is small relative to the across-seed standard deviation indicated by the error bars, and there is no abrupt degradation at any specific offset.
- **Mild and symmetric degradation.** Performance is highest at $\alpha = 1.0$, as expected, and decreases gradually and roughly symmetrically as the prior moves away from the truth. Even at $\pm 30\%$ misspecification, F1 remains within a few percentage points of its oracle value.

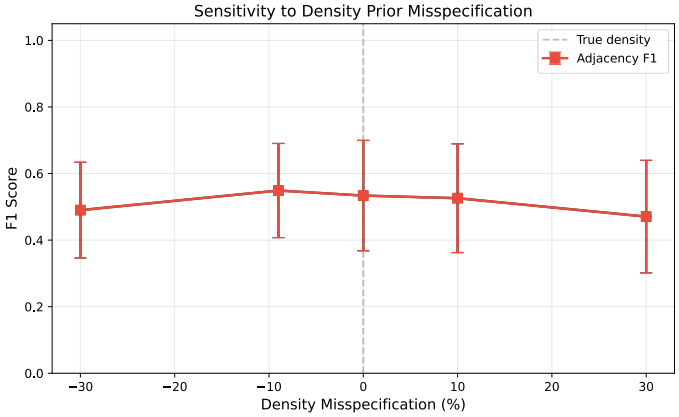

Figure 10: Sensitivity of RnR to misspecification of the density prior d. Each point aggregates 10 random seeds across all (n, extra, u) configurations; error bars show ± 1 standard deviation across seeds. The dashed vertical line marks the correct (oracle) density.

Conclusion. The density component C_o behaves as a *soft regulariser* rather than a brittle hard constraint: recovery accuracy is dominated by the edge-matching terms C_b and C_d, with the density prior shaping the cardinality of the candidate solution without locking RnR onto a specific count. Practically, this means a coarse estimate of d —- within roughly $\pm 30\%$ of the true density — is sufficient for near-optimal performance, and instance-specific tuning of the density prior is not required for real-data applications where d^* is unknown.

D RnR Algorithm Pseudocode

To improve reproducibility and address clarity concerns, we provide an explicit algorithmic description of our method. This section serves as a high-level implementation guide and complements the formal optimization framework presented earlier. Figure 2 shows the overall workflow of our method. We define the cost function as:

$$
\begin{aligned}
C(G) = \quad & \lambda_d \left(\sum_{e \in H_{dir}} w_e^+ \cdot \mathbb{I}[e \notin G] + \sum_{e \notin H_{dir}} w_e^- \cdot \mathbb{I}[e \in G_{dir}] \right) \\
+ \quad & \lambda_b \left(\sum_{e \in H_{bi}} w_e^+ \cdot \mathbb{I}[e \notin G] + \sum_{e \notin H_{bi}} w_e^- \cdot \mathbb{I}[e \in G_{bi}] \right) \\
+ \quad & \lambda_o \cdot \left| \sum_{e \in \mathcal{E}} \mathbb{I}[e \in G] - D^* \right|,
\end{aligned}
\tag{5}
$$

where H_{dir} and H_{bi} denote the sets of directed and bidirected edges in the measurement graph H, respectively. G_{dir} and G_{bi} denote the directed and bidirected edges in the candidate graph. D^* represents the expected target edge count (density). The weights $\lambda_d, \lambda_b, \lambda_o$ correspond to the hierarchical priorities defined in Section 3.3.

The weights w_e^+ and w_e^- in the cost function (Equation 5) critically influence which edges the solver prefers to keep or drop. We devise an adaptive weighting scheme based on the intuition that not all edges in the initial graph $\mathcal{H}$ are equally reliable, and similarly not all potential missing edges are equally implausible. Specifically, we leverage the strength of connection evidence (e.g., correlation or a statistical score) for each edge to set these weights: For each edge e that is present in the initial graph $\mathcal{H}$ (meaning the first-order method or data proposed this connection), we assign w_e^+ inversely proportional to the confidence or strength of e. For example, if e corresponds to a high correlation or a strong GC score between two regions, we give it a high weight—the ASP solver will incur a large cost for removing this edge, so it will likely keep it unless necessary. Conversely, if e was only weakly supported (perhaps a borderline significant link), we set w_e^+ lower, indicating that dropping this edge is not heavily penalized. This adaptivity encodes our uncertainty: strong edges should be trusted more (harder to delete), weak edges can be pruned if they conflict with other constraints.

For each potential edge e that is absent in $\mathcal{H}$ (the initial method found no link), we set w_e^- to a maximum penalty by default, reflecting a strong bias against adding completely new edges unless absolutely necessary. In other words, we assume $\mathcal{H}$ is mostly correct in saying those edges are absent, so the solver will avoid introducing a new connection e because it would incur a high cost. However, there is one exception: if we have domain knowledge or secondary evidence that a certain missing edge might actually exist (e.g., known anatomical connection between those regions), we could lower w_e^- for that edge. In general, though, w_e^- is large to favor sparse solutions that do not introduce unsupported edges.

This weighting strategy was implemented by extracting pairwise correlation magnitudes from the time-series data for all region pairs. For instance, in our fMRI simulations, we computed the Pearson correlation for each pair of nodes; if an edge was present in $\mathcal{H}$, we mapped its correlation (or other score) to a weight in [0,1] range and then scaled it. Edges not in $\mathcal{H}$ were given a uniform high weight (effectively a large constant). The ASP solver thus solves a weighted min-change problem, where changes to highly credible edges are very costly, and changes to dubious edges are cheaper.

This adaptive weighting greatly improved the accuracy of the recovered graphs. It naturally handles the uncertainty in fMRI: for example, if two regions had a very low measured correlation yet some method erroneously connected them, RnR will likely remove that edge (low penalty to drop) in favor of satisfying structural constraints. On the other hand, if two regions are strongly correlated but the initial method missed the connection (say due to a too-conservative threshold), RnR will still be unlikely to add it because of the high default cost for new edges. However, if adding that edge is crucial to satisfy the DAG or density constraint, then the solver might do so, incurring the cost but yielding a more consistent graph. In summary, adaptive weights allow RnR to "decide" which edges to trust and which to doubt, rather than treating all edges uniformly.

D.1 Theoretical Guarantees of the Optimization Procedure

While the presence of measurement noise and undersampling precludes a theoretical guarantee of statistical identifiability, ensuring the recovery of the unique ground-truth graph $\mathcal{G}^*$, our ASP-based framework provides rigorous algorithmic guarantees regarding the optimization process. Unlike

heuristic search methods (e.g., greedy hill-climbing) that may converge to local minima, the branch-and-bound optimization strategy employed by the solver guarantees global optimality for the defined cost function $C(G)$ [14]. Furthermore, the retrieval of the candidate set is exhaustive: by leveraging the completeness of stable model semantics ([25]), RnR guarantees the identification of the entire equivalence class of graphs satisfying the condition $C(G) \leq \Delta$, rather than a stochastic sample of high-scoring candidates. This ensures that the solution set captures the full extent of structural uncertainty compatible with the specified constraints.

Algorithm 1 RnR Causal Graph Recovery Algorithm

Require: Initial graph H (separated into directed H_{dir} and bidirected H_{bi} sets), edge weights w^+, w^-, target density D^*, ASP solver

Ensure: Final causal graph G^* or set of plausible graphs $\{G_1, ..., G_k\}$

/* Encode ASP Optimization Problem */

Represent all candidate edges as ASP decision variables.

Encode structural constraints:

- Ensure no directed cycles between strongly connected components (DAG constraint).
- Constrain solution space to biologically plausible density limits around D^*.

Define hierarchical cost function (Eq. 5): $C(G) = \lambda_d C_{\text{dir}} + \lambda_b C_{\text{bi}} + \lambda_o C_{\text{density}}$

If H is from another method (meta-solver mode):

- For each mutual cycle $X \to Y, Y \to X$ in H_{dir}, add a bidirected candidate $X \leftrightarrow Y$, keep both directed edges, and give the directed pair a low penalty weight (Section 3.4). Skip this step if H_{bi} is non-empty.

/* Solve via ASP */

Use Clingo in `-opt-mode=optN` to retrieve all near-optimal solutions satisfying constraints.

Collect output graphs $S = \{G_1, G_2, ..., G_k\}$.

/* Post-processing */

if $k > 1$ **then**

  Identify edge agreement across solutions (robustness).

  Optionally rank or filter graphs by validation data or consensus.

end if

Return G^* as the best or consensus graph from S.

D.2 Stage-by-Stage Ablation of Lexicographic Optimization

To verify that each stage of the lexicographic optimization performs its intended role, we modified the ASP solver to selectively disable optimization stages by omitting their weak constraints. Figure 11 shows three panels. Left: Average orientation F1 across samples improves monotonically across stages ($0.211 \to 0.350 \to 0.488$), confirming that each additional optimization criterion contributes to edge accuracy. Center: Density deviation from the ground truth is near zero when density is the sole objective (0.000) and increases slightly as competing objectives are introduced (0.012, 0.057), reflecting the expected trade-off in multi-objective optimization. Right: The solution set size decreases dramatically, from $\sim 168,920$ graphs under density-only to $\sim 1,752$ with bidirected constraints and finally to ~ 23 with the full pipeline, a reduction of nearly four orders of magnitude.

For this experiment, we use random ring graphs of size 4 and 5 nodes at target densities $d \in \{0.2, 0.3\}$, simulating stable VAR data (5,000 samples, Gaussian noise σ=0.1) undersampled at $u \in \{2, 3\}$, with PCMCI (`tau_max`=2, `ParCorr`) feeding RnR; each of the 40 instances per stage runs with a 2-hour wall-clock cap. These results demonstrate the complementary roles of the three stages: density matching constrains global sparsity, producing a large set of structurally diverse candidates; bidirected-edge resolution eliminates the vast majority of inconsistent graphs; and directed-edge orientation refines the remaining set to a small, high-quality collection suitable for expert review.

E Edge-Level Frequency Calibration

To assess whether solution-set agreement serves as a reliable per-edge confidence measure, we computed the frequency of each directed edge across all near-optimal solutions returned by RnR and measured precision (fraction of edges that are true positives in the ground truth) within each

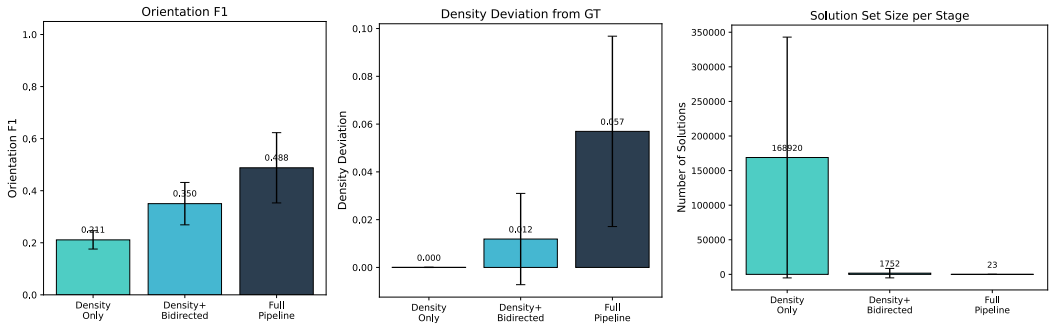

Figure 11: Stage-by-stage ablation of lexicographic optimization, average across samples. Orientation F1 (left), density deviation from ground truth (center), and solution set size (right) across three optimization stages: density only, density + bidirected, and full pipeline. Each successive stage improves F1 ($0.211 \rightarrow 0.350 \rightarrow 0.488$) while reducing the solution set by nearly four orders of magnitude ($168,920 \rightarrow 1,752 \rightarrow 23$). Error bars indicate standard deviation across runs.

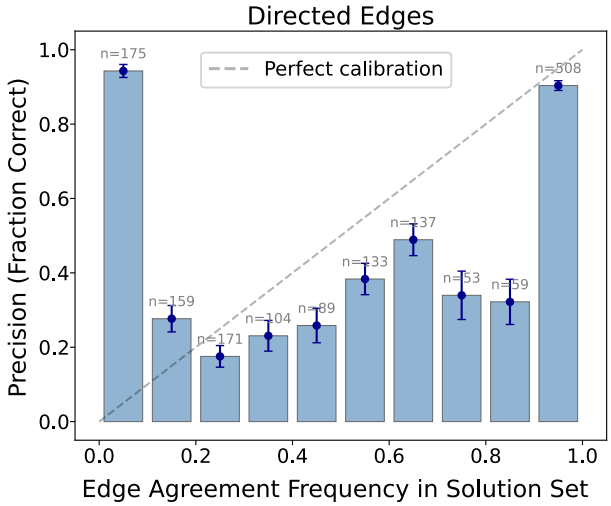

Figure 12: Precision (fraction of edges that are true positives) versus edge agreement frequency across the near-optimal solutions, computed over 100 simulation runs. The dashed diagonal marks perfect calibration. The first bin (frequency 0.0–0.1, $n = 175$) consists mostly of ground-truth edges that appear in no solution, so its height reflects missed edges rather than precision. Near-unanimous edges (frequency 0.9–1.0) reach precision 0.90 ($n = 508$).

frequency bin. The analysis spans 100 simulation runs (5 Sanchez-Romero networks, undersampling rates u=2,3, 10 batches per configuration).

Figure 12 shows precision as a function of edge agreement frequency. The first bin (frequency 0.0–0.1) contains ground-truth edges absent from all solutions, reflecting the solver's recall ceiling rather than a confidence signal, and is excluded from the calibration interpretation. Agreement is informative mainly near unanimity: edges present in at least 90% of solutions are correct 90% of the time ($n = 508$), whereas across the bins from 0.1 to 0.9 precision stays between 0.18 and 0.49 and does not rise monotonically.

Edge frequency is therefore a coarse filter rather than a calibrated probability: near-unanimous edges are reliable, and edges below that level should be treated as uncertain. The design objective of RnR is not to recover a single point estimate but to return a small set of near-optimal solutions for expert consideration, and edge frequency gives a simple way to separate the edges the set agrees on from those it does not.

F Brief Introduction to `clingo` and Answer Set Programming (ASP)

This section provides a brief technical overview of Answer Set Programming (ASP) and the solver `clingo`. ASP is a declarative problem-solving paradigm tailored for difficult combinatorial search and optimization problems. Unlike imperative programming, where the user specifies *how* to compute a solution, in ASP the user specifies *what* constitutes a problem and its valid solutions.

The tool `clingo` is a monolithic system that integrates two primary processes:

1. **Grounding (Gringo):** Instantiates a logical program containing variables into a variable-free (propositional) format.
2. **Solving (Clasp):** Performs conflict-driven nogood learning (CDNL) to find stable models (answer sets) that satisfy the grounded logic program.

A typical ASP encoding follows a "Generate-Define-Test" methodology. Below, we outline the core syntactic constructs of `clingo` using the classic *Graph Coloring Problem* as an illustrative example.

F.1 Facts: The Problem Instance

Facts represent the unconditional truths of the specific problem instance. These are atoms (predicates) that are true at the start of the solving process. In a graph context, facts usually define the topology, such as nodes and edges.

Listing 1: Defining a graph instance and available colors.

```
node(1..4).          % Shorthand for node(1), node(2), etc.
edge(1,2). edge(2,3). edge(3,4). edge(4,1). edge(1,3).
color(red). color(blue). color(green).
```

In Listing 1, `node(1..4)` utilizes `clingo`'s range syntax to generate facts for nodes 1 through 4 efficiently.

F.2 Rules: Logical Derivation

Normal rules are used to derive new information based on existing facts or other derived atoms. A rule is structured as `Head :- Body.` and follows the logic: "If the Body is true, then the Head must be true." Variables in ASP are denoted by uppercase letters and are universally quantified.

For undirected graphs, we might want to ensure that if an edge exists from X to Y, the system treats it as symmetric:

Listing 2: Deriving symmetric edges.

```
% If there is an edge from X to Y, imply an edge from Y to X
edge(Y, X) :- edge(X, Y).
```

During the grounding phase, `clingo` replaces variables X and Y with all valid constants from the domain (e.g., numbers 1 to 4) to create propositional clauses.

F.3 Choice Rules: The Generator

To solve combinatorial problems, the solver must explore the search space by making choices. This is achieved via *choice rules* (or cardinality constraints). A choice rule allows the solver to select a subset of atoms to include in a candidate model.

The syntax `{ A } = N` asserts that exactly N atoms from set A must be true.

Listing 3: Generating color assignments.

```
% For every node N, choose exactly one color C from the defined colors.
1 { assign(N, C) : color(C) } 1 :- node(N).
```

In Listing 3, the construct `1 { ... } 1` enforces a strict constraint: for each specific node N, the solver must pick a minimum of 1 and a maximum of 1 color assignment. This generates the search space of all possible valid and invalid colorings.

F.4 Integrity Constraints: The Tester

While choice rules generate the potential search space, *integrity constraints* prune this space by eliminating invalid solutions. An integrity constraint is a rule with an empty head, written as `:- Body.` It signifies: "It must not be the case that the Body is true."

If the body of an integrity constraint is satisfied by a candidate model, that model is discarded.

Listing 4: Eliminating invalid graph colorings.

```
% It is strictly forbidden for two connected nodes to share a color.
:- edge(X, Y), assign(X, C), assign(Y, C).
```

Listing 4 ensures that if two nodes X and Y share an edge, they cannot be assigned the same color C. If the solver attempts a model where connected nodes are both red, this constraint triggers, and the solver backtracks.

F.5 Solving and Answer Sets

When `clingo` executes the program combining the facts, generator, and constraints, it outputs *Answer Sets*. Each Answer Set is a minimal stable model representing a valid solution to the problem.

For the example above, a resulting answer set might look like:

```
assign(1,red) assign(2,blue) assign(3,green) assign(4,blue)
...
```

If no model satisfies all constraints (e.g., trying to color a fully connected graph of 4 nodes with only 3 colors), `clingo` will return `UNSATISFIABLE`.

F.6 Advanced Features: Optimization

`clingo` also supports optimization statements to find the "best" models among the valid ones. For example, if we assigned a cost to using specific colors, we could instruct the solver to minimize the total cost:

```
% Minimize the sum of Cost for all chosen assignments
#minimize { Cost,N : assign(N, C), cost(C, Cost) }.
```

This feature allows `clingo` to be used not just for satisfiability tasks, but for complex optimization problems common in machine learning pipelines and logic-based reasoning.

G fMRI Data and Preprocessing

We used resting-state fMRI from the FBIRN cohort [23]: 150 healthy subjects, TR $= 2$ s, 140 volumes per subject. Volumes were preprocessed with the standard NeuroMark pipeline and decomposed by spatially constrained ICA against the NeuroMark_fMRI_1.0 template [11], which yields 53 intrinsic connectivity networks in seven domains. Following Sridharan et al. [44], six networks were fixed a priori by anatomy: rPPC (ICN 25), rFIC (ICN 26) and rDLPFC (ICN 35) from the cognitive-control domain, and ACC (ICN 44), PCC (ICN 45) and VMPFC (ICN 46) from the default-mode domain.

For GCM, component time courses were de-meaned and z-scored. Following Roebroeck et al. [36], the bivariate VAR order was selected per pair over $p \in [1, 8]$ by the Schwarz criterion, and Geweke's $F_{x\to y} - F_{y\to x}$ was tested against 200 block-bootstrap surrogates (block length 16), with an edge declared at $p < 0.05$. For RnR, the same six time courses were passed to PCMCI (ParCorr, $\tau_{\max} = 1$) to obtain $\mathcal{H}$, which RnR refined under SCC constraints derived from the NeuroMark domain structure. Frequencies in Table 2 pool over subjects for GCM, and over subjects and near-optimal solutions for RnR.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We include experiments and show results to support our claims in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the inherent limitations of the research question and how in some cases, the correct answer is really not known. We also discuss the limitation of our method in terms of a scale and discuss how we will address them in future works.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: our paper does not include any theoretical results or theorems.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we include all the hyperparameters needed to reproduce the results and the pseudocode in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: https://github.com/neuroneural/gunfolds

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: yes, we have included all the hyperparameters used in this method.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: yes, all of our results and figures include error bars. Since our method deals with discrete answers, while reporting all the experiments, we ran in the same setting 10, 50, or in some cases 100 times with random seeds. We only report the summary statistics of those results to account for any outliers.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: all the experiments share the same resources and setup. We explain our hardware here in Appendix A.3

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both in the last section of the paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All code and data we use are cited. FBIRN Phase III data [23] are used under the BIRN Data Use Agreement: no re-identification, the FBIRN grants are acknowledged, and the data are not redistributed. Brain regions come from the NeuroMark_fMRI_1.0 template [11], distributed with the GIFT toolbox (https://github.com/trendscenter/gift); the repository states no license, and its creators include an author of this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.