# MULTI-OMIC CAUSAL DISCOVERY USING GENOTYPES AND GENE EXPRESSION

**Stephen Asiedu**
**David Watson**
Department of Informatics
King's College London
{stephen.asiedu, david.watson}@kcl.ac.uk

## ABSTRACT

Causal discovery in multi-omic datasets is crucial for understanding the bigger picture of gene regulatory mechanisms but remains challenging due to high dimensionality, differentiation of direct from indirect relationships, and hidden confounders. We introduce GENESIS (GEne Network inference from Expression SIgnals and SNPs), a constraint-based algorithm that leverages the natural causal precedence of genotypes to infer ancestral relationships in transcriptomic data. Unlike traditional causal discovery methods that start with a fully connected graph, GENESIS initializes an empty ancestrality matrix and iteratively populates it with direct, indirect or non-causal relationships using a series of provably sound marginal and conditional independence tests. By integrating genotypes as fixed causal anchors, GENESIS provides a principled "head start" to classical causal discovery algorithms, restricting the search space to biologically plausible edges. We test GENESIS on synthetic and real-world genomic datasets. This framework offers a powerful avenue for uncovering causal pathways in complex traits, with promising applications to functional genomics, drug discovery, and precision medicine.

## 1 INTRODUCTION

In recent years, high-throughput technologies have generated vast volumes of multi-omic data across different layers such as transcriptomics, proteomics, and metabolomics (Qiao et al., 2024; Reuter et al., 2015; Manel et al., 2016). While this wealth of information holds immense potential for advancing our understanding of biological systems (Abu-Elmagd et al., 2022; Bourne et al., 2015), it simultaneously poses formidable analytical challenges, particularly in the realm of complexity and high-dimensionality (Hu et al., 2018). Although considerable progress has been made in applying causal discovery methods within individual omics layers, these approaches fail to capture the intricate interplay between the different molecular networks. Integrating information across multiple omics layers promises to reveal more detailed explanations that would remain obscured if we solely rely on single-omics analyses (Danchin et al., 2007; Veenstra, 2012; Neale & Wheeler, 2019). Adopting a multi-omic causal discovery approach is essential to advance our biological understanding and design more targeted therapeutic interventions (He et al., 2017; Mohammadi-Shemirani et al., 2023).

A major challenge in modern genomics is to infer the gene regulatory networks (GRNs) that dictate cellular behavior (Karlebach & Shamir, 2008). A clear and precise understanding of GRNs can illuminate the pathways that lead to complex traits and diseases. However, the underlying data is inherently high-dimensional, posing major statistical and computational challenges. One promising strategy, which we build on below, is the usage of cis-expression quantitative trait loci (cis-eQTLs) (Michaelson et al., 2009), which leverage two key biological facts: (1) that genetic variation precedes transcriptomic variation; and (2) that the influence of a genetic variant on a target gene decreases as a function of spatial proximity. This targeted approach boosts statistical power by exploiting prior knowledge and reducing the search space.

In this context, we introduce GEne NEtwork inference from expression SIgnals and Single nucleotide polymorphisms (GENESIS), an algorithm to integrate genotype and transcriptomic data to reconstruct directed GRNs. Our method harnesses the natural variability of SNPs to distinguish between direct and indirect gene regulatory effects by following the inference rules proposed by Magliacane et al. (2016). We implement a two-step hypothesis testing framework to identify marginal SNP-gene associations within cis-windows and filter out indirect effects via conditional independence tests. We use parametric plug-ins based on linear modeling assumptions that are widely used in bioinformatics (Smyth, 2004; Love et al., 2014), ensuring the efficiency of our method. GENESIS is provably sound, delivering a partially oriented ancestrality matrix in polynomial time that can lead to major speedups when used as a preprocessing step for classical causal discovery methods like the PC algorithm (Spirtes et al., 2001). We illustrate our method on simulated and real-world data, where it compares favorably to the state of the art.

## 2 BACKGROUND

We use upper case italics $X$ to denote random variables or sets thereof. Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ be a graph with nodes $\mathcal{V}$ that represent variables and directed edges $\mathcal{E} : \mathcal{V} \times \mathcal{V}$ that denote causal relationships between them. We focus in particular on directed acyclic graphs (DAGs), as is common in causal discovery (Spirtes et al., 2001; Peters et al., 2017). In the context of this study, nodes may represent SNPs (background variables $Z$) or genes (foreground variables $X$). We aim to draw inferences about the relationships between the latter by exploiting signals from the former. Specifically, our goal is to infer as much as possible about the subgraph $\mathcal{G}_X = \langle X, \mathcal{E}_X \rangle$, with edge set $\mathcal{E}_X : \{Z \cup X\} \times X$ including all directed arrows into foreground variables.

We use kinship terms to describe relationships between nodes, with $Pa(\cdot), Ch(\cdot), An(\cdot), De(\cdot)$ representing the parents, children, ancestors, and descendants (respectively) of a given node set. If $X \in An(Y)$ (or, equivalently, $Y \in De(X)$), we write $X \prec Y$ (equivalently, $Y \succ X$). If $X \notin De(Y)$, we write $X \preceq Y$. We write $X \sim Y$ when neither variable is an ancestor of the other. Ancestry graphs impose a strict partial order on nodes, characterized by the following properties: (1) *irreflexive*: $X \prec X \Rightarrow$ `FALSE`; (2) *asymmetric*: $X \prec Y \Rightarrow Y \not\prec X$; and (3) *transitive*: $X \prec Y \,\&\, Y \prec Z \Rightarrow X \prec Z$.

We use standard probabilistic definitions of independence, writing $X \perp\!\!\!\perp Y \mid Z$ to indicate that variable sets $X$ and $Y$ have no mutual information after conditioning on the (potentially empty) variable set $Z$. We assume that distributions are Markov and faithful to the underlying graph $\mathcal{G}$, in which case conditional independence claims are equivalent to $d$-separation statements (Pearl, 2009).

Building on the work of Claassen & Heskes (2012) and Watson & Silva (2022), we introduce the concept of (de)activators.

**Definition 1** (Deactivator)**.** *A variable $W$ is a deactivator of the relationship between $X$ and $Y$ given $Z$ if (a) $X \not\perp\!\!\!\perp Y \mid Z$; and (b) $X \perp\!\!\!\perp Y \mid Z \cup W$. In this case, we write $X \perp\!\!\!\perp Y \mid Z \cup [W]$.*

A deactivator is a single variable that, when added to the conditioning set $Z$, is sufficient to block all otherwise open paths linking $X$ and $Y$.

**Definition 2** (Activator)**.** *A variable $W$ is an activator of the relationship between $X$ and $Y$ given $Z$ if (a) $X \perp\!\!\!\perp Y \mid Z$; and (b) $X \not\perp\!\!\!\perp Y \mid Z \cup W$. In this case, we write $X \not\perp\!\!\!\perp Y \mid Z \cup [W]$.*

An activator is a single variable that, when added to the conditioning set $Z$, is sufficient to open an otherwise blocked path between $X$ and $Y$.

## 3 METHOD AND ALGORITHM

In this section, we present the oracle version of GENESIS, which is designed to infer causal relationships among a set of variables through a careful series of conditional independence (CI) queries. Unlike traditional methods that begin with a fully connected graph and progressively remove edges (e.g., the PC algorithm), GENESIS-ORACLE is initialized with an empty ancestrality matrix $\mathbf{M}$ and incrementally adds causal information based on the results of calls to the oracle. Inputs include a background set $Z$ (e.g., SNPs or other exogenous data) and a foreground set $X$ (e.g., gene expres-

sion data), while the output is the ancestrality matrix $\mathbf{M}$ in which each element $\mathbf{M}_{ij}$ encodes the causal relationship (if decidable) between pairs of foreground variables $X_i, X_j$.

GENESIS relies on three simple inference rules, variants of which have appeared in several causal discovery methods (Claassen & Heskes, 2012; Entner et al., 2013; Watson & Silva, 2022). Let $A$ and $\{X, Y\}$ be two sets of nodes in the graph $\mathcal{G}$ where $A \preceq \{X, Y\}$, and let $A_{\setminus W} := A \setminus \{W\}$ for some node $W \in A$. Our first rule detects causal pathways via deactivation patterns:

(R1) If $\exists W \in A : W \perp\!\!\!\perp Y \mid A_{\setminus W} \cup [X]$, then $X \prec Y$.

This is only possible when $W$ is a mediator on the path from $X$ to $Y$. Our second rule rejects causal pathways via activation patterns:

(R2) If $\exists W \in A : W \not\perp\!\!\!\perp X \mid A_{\setminus W} \cup [Y]$, then $X \preceq Y$.

This is only possible when $Y$ is a (descendant of a) collider on the path from $W$ to $X$ and is not a non-collider on any other path active under $A_{\setminus W}$. Finally, our third rule establishes causal independence via $d$-separation:

(R3) If $X \perp\!\!\!\perp Y \mid A$, then $X \sim Y$.

See Alg. 1 for a summary of the oracle procedure. We use the grow-shrink algorithm to infer the Markov blanket of each $X$, as this method is efficient and sound (Margaritis & Thrun, 1999). Alternatives are possible in practice, e.g. IAMB (Tsamardinos et al., 2003). Keeping a running tab of each variable's Markov blanket can lead to major speedups over alternative procedures such as the confounder blanket learner (Watson & Silva, 2022), which must cycle through vast conditioning sets for each pair of foreground variables.

Once Markov blankets are initialized, the basic procedure is to cycle through all variable pairs. By (R3), we conclude that any foreground variables that are $d$-separated by their combined Markov blanket $S$ must be causally unconnected. Next, we loop through the elements of $S$, applying (R2) and (R3) to test for patterns of (de)activation that can help orient causal relations within $X$. Once we have done this for all pairs, we use a set of closure rules (see Alg. 2 in Appx. A) that exploits the strict partial order on $\mathcal{G}_X$ to potentially draw some extra inferences about $\mathbf{M}$. Finally, we update the Markov blanket for each foreground variable in case any newly inferred non-descendants might enter in. The algorithm converges either when the ancestral graph is fully oriented or a complete pass fails to draw any new inferences.

We establish some basic properties of this algorithm. (For proofs, see Appx. B.)

**Theorem 1** (Soundness)**.** *All inferences returned by* GENESIS-ORACLE *hold in the true* $\mathcal{G}_X$*. Moreover, if* $\mathbf{M}_{ij} = i \prec j$*, then the set of combined Markov blankets* $S = MB(X_i) \cup MB(X_j)$ *is a valid adjustment set for* $(X_i, X_j)$*.*

This result follows from the soundness of the inference rules themselves, which has been previously established by numerous authors.

**Theorem 2** (Complexity)**.** *Let* $d_Z, d_X$ *be the dimensionality of the background variables* $Z$ *and foreground variables* $X$*, respectively. Then the complexity of* GENESIS-ORACLE *is* $\mathcal{O}(d_Z d_X^2)$*.*

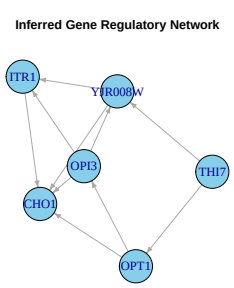
This result holds regardless of graph density. In sparse settings, runtime can be highly efficient due to the relatively low cardinality of $S$.

In summary, our method builds its causal structure from the ground up by gradually adding evidence-based edges. This forward-construction approach not only avoids the exhaustive edge assignment typical of fully connected initializations but also supplies a robust starting point for downstream causal discovery algorithms, such as FCI or the PC algorithm. By constraining further searches to only those edges consistent with the inferred ancestry matrix, our method significantly improves both speed and reliability, particularly in high-dimensional genomic settings where disentangling direct and indirect regulatory effects is critical.

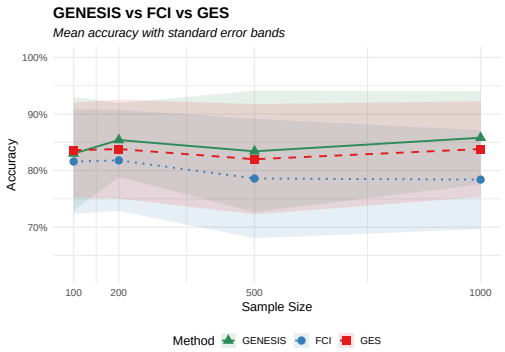
**Algorithm 1** GENESIS-ORACLE

**Input**: Background variables $Z$, foreground variables $X$
**Output**: Ancestrality matrix $\mathbf{M}$

Initialize: `converged` $\leftarrow$ `FALSE`, $\mathbf{M} \leftarrow$ [NA]
**for all** $X_i \in X$ **do**
  $MB(X_i) \leftarrow$ `GrowShrink`$(X_i, Z)$
**end for**
**while not** `converged` **do**
  `converged` $\leftarrow$ `TRUE`
  **for all** $(X_i, X_j) \in X$ s.t. $i < j, \mathbf{M}_{ij} =$ NA **do**
    $S \leftarrow MB(X_i) \cup MB(X_j)$
    **if** $X_i \perp\!\!\!\perp X_j \mid S$ **then**
      $\mathbf{M}_{ij} \leftarrow i \sim j$, `converged` $\leftarrow$ `FALSE`
    **else**
      **for** $W \in S$ **do**
        **if** $W \perp\!\!\!\perp X_j \mid S_{\setminus W} \cup [X_i]$ **then**
          $\mathbf{M}_{ij} \leftarrow i \prec j$, `converged` $\leftarrow$ `FALSE`
        **else if** $W \perp\!\!\!\perp X_i \mid S_{\setminus W} \cup [X_j]$ **then**
          $\mathbf{M}_{ij} \leftarrow j \prec i$, `converged` $\leftarrow$ `FALSE`
        **else if** $W \not\perp\!\!\!\perp X_j \mid S_{\setminus W} \cup [X_i]$ **then**
          $\mathbf{M}_{ij} \leftarrow \mathbf{M}_{ij} \wedge j \preceq i$, `converged` $\leftarrow$ `FALSE`
        **else if** $W \not\perp\!\!\!\perp X_i \mid S_{\setminus W} \cup [X_j]$ **then**
          $\mathbf{M}_{ij} \leftarrow \mathbf{M}_{ij} \wedge i \preceq j$, `converged` $\leftarrow$ `FALSE`
        **end if**
      **end for**
    **end if**
  **end for**
  $\mathbf{M} \leftarrow$ `Closure`$(\mathbf{M})$
  **for all** $X_i \in X$ **do**
    $A_i \leftarrow MB(X_i) \cup \{W : W \preceq_{\mathbf{M}} X_i\}$
    $MB(X_i) \leftarrow$ `GrowShrink`$(X_i, A_i)$
  **end for**
**end while**




(a) Gene regulatory network for the phosphocoline subnetwork in *Saccharomyces cerevisiae* predicted by GENESIS




(b) A multivariate comparison of GENESIS with FCI (constraint-based algorithm) and GES (Score-based algorithm)

Figure 1: Results on real world and simulated data

## 4 RESULTS AND CONCLUSION

We analyzed the yeast dataset provided by the TRIGGER package (Chen et al., 2007), which comprises 112 recombinant haploid segregant strains derived from a cross between two haploid parental

strains of Saccharomyces cerevisiae (BY and RM) (Brem & Kruglyak, 2005). This dataset includes genome-wide expression profiles for 6,216 genes and genotypic information from 3,244 single-nucleotide polymorphism (SNP) markers. One key strategy we employed was the analysis of cis-expression quantitative trait loci (cis-eQTLs), which exploits the physical proximity between SNPs and their associated genes. By limiting our search to a 5 kilobase genomic window, which is a standard practice in cis-eQTL mapping, we prioritized genetic variants most likely to exert direct regulatory effects. Our investigation focused exclusively on genes previously identified in the literature as components of the phosphocholine subnetwork, which governs the biosynthesis of phosphatidylcholine through the Kennedy pathway. This pathway is essential for maintaining membrane integrity and supporting critical cellular processes in yeast (Henneberry et al., 2001). To assist in identifying our iterative proximal ancestor sets, we utilized partial correlation tests within the GENESIS framework. The resulting regulatory network, shown in Figure 1a, adheres to an acyclicity constraint and captures ancestral relationships that are consistent with those reported in previous work on the yeast phosphocholine network (Chen et al., 2019).

To evaluate the performance of GENESIS in a controlled multivariate setting, we conducted a series of simulations using randomly generated directed acyclic graphs (DAGs) with both background $Z$ and foreground $X$ variables with the goal of inferring ancestral relationships $X$. The DAGs were randomly generated with varying sample size ranging between 100 and 1000 with edge densities chosen to maintain moderate sparsity. Our data were simulated from these structures using a linear Gaussian model with mixed noise. For each sample size, we compared the structural recovery accuracy of GENESIS against two well-established causal discovery methods: the Fast Causal Inference (FCI) (Spirtes et al., 2001) algorithm and Greedy Equivalence Search (GES) (Chickering, 2002). GENESIS was run with a conditional independence threshold of $\alpha = 0.05$ while FCI and GES were configured using default parameters. Performance was quantified as the mean adjacency matrix accuracy over 20 replicates, computed as the element-wise match between the inferred and true adjacency matrices among target variables. Across the sample sizes, GENESIS consistently performed better than the benchmark methods.

While GENESIS demonstrates strong empirical performance in both simulated and biological settings, its reliance on iterative heuristics and sensitivity to initialization highlight the need for further theoretical analysis and optimization. Future work may explore extensions to nonlinear models and formal guarantees for convergence, completeness, and identifiability under broader conditions.

## REFERENCES

M. Abu-Elmagd, Mourad Assidi, A. Alrefaei, and Ahmed Rebai. Editorial: Advances in genomic and genetic tools, and their applications for understanding embryonic development and human diseases. *Frontiers in Cell and Developmental Biology*, 10, 2022. doi: 10.3389/fcell.2022.1016400.

Philip E Bourne, Jon R Lorsch, and Eric D Green. Perspective: Sustaining the big-data ecosystem. *Nature*, 527(7576):S16–S17, 2015.

Rachel B Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences*, 102(5):1572–1577, 2005.

Chen Chen, Dabao Zhang, Tony R Hazbun, and Min Zhang. Inferring gene regulatory networks from a population of yeast segregants. *Scientific reports*, 9(1):1197, 2019.

Lin S Chen, Frank Emmert-Streib, and John D Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome biology*, 8:1–13, 2007.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Tom Claassen and Tom Heskes. A logical characterization of constraint-based causal discovery. *arXiv preprint arXiv:1202.3711*, 2012.

A. Danchin, Gang Fang, and Stanislas Noria. The extant core bacterial proteome is an archive of the origin of life. *PROTEOMICS*, 7, 2007. doi: 10.1002/PMIC.200600442.

Doris Entner, Patrik Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Artificial intelligence and statistics*, pp. 256–264. PMLR, 2013.

K. He, Dongliang Ge, and Max M He. Big data analytics for genomic medicine. *International Journal of Molecular Sciences*, 18, 2017. doi: 10.3390/ijms18020412.

Annette L Henneberry, Thomas A Lagace, Neale D Ridgway, and Christopher R McMaster. Phosphatidylcholine synthesis influences the diacylglycerol homeostasis required for sec14p-dependent golgi function and cell growth. *Molecular biology of the cell*, 12(3):511–520, 2001.

Pengfei Hu, Rong Jiao, Li Jin, and Momiao Xiong. Application of causal inference to genomic analysis: Advances in methodology. *Frontiers in Genetics*, 9, 2018. URL `https://api.semanticscholar.org/CorpusID:49652600`.

Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, 9(10):770–780, 2008.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15:1–21, 2014.

Sara Magliacane, Tom Claassen, and Joris M Mooij. Ancestral causal inference. *Advances in Neural Information Processing Systems*, 29, 2016.

Stéphanie Manel, Charles Perrier, M. Pratlong, Laurent Abi-Rached, J. Paganini, Pierre Pontarotti, and Didier Aurelle. Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Molecular Ecology*, 25, 2016. doi: 10.1111/mec.13468.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. *Advances in neural information processing systems*, 12, 1999.

Jacob J Michaelson, Salvatore Loguercio, and Andreas Beyer. Detection and interpretation of expression quantitative trait loci (eqtl). *Methods*, 48(3):265–276, 2009.

Pedrum Mohammadi-Shemirani, Tushar Sood, and Guillaume Paré. From 'omics to multi-omics technologies: the discovery of novel causal mediators. *Current atherosclerosis reports*, 25(2):55–65, 2023.

D. Neale and N. Wheeler. Gene expression and the transcriptome. *The Conifers: Genomes, Variation and Evolution*, 2019. doi: 10.1007/978-3-319-46807-5_6.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Yi Qiao, Tianguang Cheng, Zikun Miao, Yue Cui, and Jing Tu. Recent innovations and technical advances in high-throughput parallel single-cell whole-genome sequencing methods. *Small Methods*, pp. 2400789, 2024.

Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597, 2015.

Gordon K Smyth. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS*, volume 2, pp. 376–81, 2003.

T. Veenstra. Metabolomics: the final frontier? *Genome Medicine*, 4:40 – 40, 2012. doi: 10.1186/gm339.

David S Watson and Ricardo Silva. Causal discovery under a confounder blanket. In *Uncertainty in Artificial Intelligence*, pp. 2096–2106. PMLR, 2022.

## A CLOSURE

Below is the pseudocode for the closure described in Algorithm 1.

**Algorithm 2** CLOSURE

**Input:** Ancestrality matrix $\mathbf{M}$
**Output:** Updated ancestrality matrix $\mathbf{M}$

**for** $i, j \in \{1, \ldots, d_X\}$ such that $i > j$ **do**
  **if** $i \preceq_{\mathbf{M}} j \wedge i \succeq_{\mathbf{M}} j \vee i \sim_{\mathbf{M}} j$ **then**
    $\mathbf{M}_{ij} \leftarrow i \sim j$
  **else if** $i \prec_{\mathbf{M}} j$ **then**
    $\mathbf{M}_{ij} \leftarrow i \prec j$
  **else if** $j \prec_{\mathbf{M}} i$ **then**
    $\mathbf{M}_{ij} \leftarrow j \prec i$
  **end if**
**end for**
`converged` $\leftarrow$ `FALSE`
**while not** `converged` **do**
  `converged` $\leftarrow$ `TRUE`
  **for** $i, j, k \in \{1, \ldots, d_X\}$ such that $i \neq j \neq k, i > k$ **do**
    **if** $i \prec_{\mathbf{M}} j \prec_{\mathbf{M}} k \wedge \mathbf{M}_{ik} \neq i \prec k$ **then**
      $\mathbf{M}_{ik} \leftarrow i \prec k$, `converged` $\leftarrow$ `FALSE`
    **else if** $k \prec_{\mathbf{M}} j \prec_{\mathbf{M}} i \wedge \mathbf{M}_{ik} \neq k \prec i$ **then**
      $\mathbf{M}_{ik} \leftarrow k \prec i$, `converged` $\leftarrow$ `FALSE`
    **end if**
  **end for**
**end while**

## B PROOFS

**Theorem 1 (Soundness)**

*Proof.* By construction, GENESIS-ORACLE only applies the three sound rules (R1), (R2) and (R3) to the union of the Markov blankets of $X_i$ and $X_j$. However, we feed the `GrowShrink` algorithm with only known non-descendants and hence the guarantee that the union of Markov blankets will be non-descendants themselves. We then apply closure under transitivity and asymmetry as seen in Alg 2. This means the soundness of our oracle depends on the soundness of the rules themselves. (R1) and (R2) follows from a direct application of Lemma 1 from Magliacane et al. (2016) while (R3) is the direct application of faithfulness since we are limited to non-descendants.

For us to arrive at $M_{ij} = i \prec j$, we must use (R1) with some $S = MB(X_i) \cup MB(X_j)$ to detect a minimal deactivation of the form $W \perp\!\!\!\perp X_j \mid S_{\setminus W} \cup [X_i]$ for some $W \in S$ as proved by Entner et al. (2013) (where $S$ is limited to a set of non-descendants). As stated earlier, the union of Markov blankets $S$ of $(X_i, X_j)$ will solely contain non-descendants as we start with $Z$ which is biologically a well establish non-descendant of $X_i \in X$ and $W \in S$. Since this satisfies the assumption of Entner et al. (2013), we conclude, using the same argument, that $S = MB(X_i) \cup MB(X_j)$ is a valid adjustment set for $(X_i, X_j)$. □

**Theorem 2 (Complexity)**

*Proof.* From the GENESIS-ORACLE 1, the initialization of the Markov blanket will cost $\mathcal{O}(d_X d_Z)$ as each foreground variables requires a pass through the background variables. Each unordered pair $(X_i, X_j)$, with $i < j$, is resolved at most once per execution of the `while` loop. There are $\binom{d_X}{2} = \mathcal{O}(d_X^2)$ such pairs in total for the pairwise inference loop. Initially since $S$ will contain at most $d_Z$ elements, it will cost $\mathcal{O}(d_Z)$ and $\mathcal{O}(d_Z + d_X)$ in subsequent iterations. We assume that $d_Z \gg d_X$, in which case this reduces to $\mathcal{O}(d_Z)$. In the oracle setting, the CI queries execute in constant time,

$\mathcal{O}(1)$, although practical implementations tend to scale with the size of the conditioning set. Thus the pairwise inference loop will cost $\mathcal{O}(d_Z) \times \mathcal{O}(d_X^2) = \mathcal{O}(d_Z d_X^2)$. □