Causal Discovery of Radiation Response Mechanisms in Human Cells

Ashka Shah SHAHASHKA@UCHICAGO.EDU
Department of Computer Science
University of Chicago
Chicago, IL, USA

Rick Stevens STEVENS@CS.UCHICAGO.EDU
Department of Computer Science
University of Chicago
Chicago, IL, USA

Abstract

Next-generation sequencing technologies, including RNA-sequencing, provide genome-wide measurements of gene expression and enable broad explorations of biomarkers and mechanisms underlying disease and treatment response. Bioinformatics tools for processing this data, such as differential expression analysis, are largely univariate, linear, and rely on predefined pathway knowledge annotations, which limits their ability to capture nonlinear and multivariate gene interactions. This paper explores the application of causal discovery to characterizing transcriptional responses to radiation as a function of dose rate in human cells. By jointly modeling radiation perturbations and gene expression, we learn directed gene networks that capture important regulatory relationships beyond correlation and exhibit significant enrichment of known radiation response pathways compared to baseline approaches. We find that inferred causal graphs reveal structured network features such as high in-degree housekeeping genes and high out-degree transcription factors. Further analysis suggests a hierarchical organization of stress response pathways and triggered cell death pathways. This work highlights the potential of causal discovery in healthcare settings with applications to understanding response mechanisms, identifying regulatory targets, and improving interpretation of complex genomic data. Code is available at `https://github.com/shahashka/lucid_cd`.

1. Introduction

RNA-sequencing experiments provide genome-wide insight into gene expression levels across conditions at high resolution; however, finding genes of interest from this vast data remains a difficult challenge (Trapnell et al., 2009; Love et al., 2014). Differential expression analysis is the most popular tool used to identify genes that statistically vary between perturbed and control samples as an initial step to identify important genes (Rosati et al., 2024). Subsequent pathway enrichment onto knowledge bases is used to understand the molecular mechanisms underlying the conditions analyzed. Finally, biomarkers are found by linking genes to known disease pathways. Despite its popularity, differential expression analysis combined with pathway enrichment suffers from two main drawbacks: (1) genes are identified using linear models and univariate tests; this misses coordinated and nonlinear

signals from genes that may individually be weak (2) its outputs do not directly imply any mechanistic understanding limiting discovery of novel mechanisms and pathways. To address these issues multivariate tests (Glazko and Emmert-Streib, 2009), supervised machine learning with feature ranking (Wenric and Shemirani, 2018), and graph learning algorithms (Marbach et al., 2012; Moerman et al., 2019) have been applied to identify important genes under perturbed conditions.

Recently causal discovery (algorithms for inferring cause and effect relationships represented by graphs) has emerged as a data-driven approach to mechanism discovery. Causal discovery often focuses on synthetic data because real world data violate the assumptions needed for identifiability of the true causal graph; these include the absence of confounding variables, acyclic relationships, and large sample sizes. There is a growing call for the application of these methods to more realistic and scientific settings (Brouillard et al., 2024). To this end, benchmarking frameworks with large-scale transcriptomics and CRISPR perturbation datasets have been developed; however, they largely reveal that causal discovery algorithms fail to learn regulatory relationships between genes from single-cell RNA-sequencing datasets (Chevalley et al., 2025).

In this paper, we take an alternative approach by applying causal discovery to low-sample, perturbed-control transcriptomics studies for understanding radiation response. We jointly model radiation exposure, which is viewed as a perturbation upstream of gene regulation that affects many pathways, alongside the multivariate nonlinear distribution of gene expression. We apply this framework to RNA sequencing data sampled from RPE1 cell lines exposed to chronic low-dose radiation across increasing dose rates. Such prolonged exposure is associated with elevated risks of cancer and cardiovascular disease and is known to activate cellular stress-response pathways (Shimizu et al., 2010). A complete understanding of the interplay between dose rate, cumulative dose, and cell type in the activation of low-dose radiation pathways remains an open question (Fig. 1) and has implications for people chronically exposed to low-dose radiation (e.g., space exploration, radon exposure) (Verheij and Bartelink, 2000). To address this gap, we use causal discovery to uncover relationships linking radiation exposure to gene expression, and provide mechanistic insight into how dose rate affects cellular stress-response pathways.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work applies causal discovery to infer important genes and to learn novel pathways from high-dimensional biological data with limited samples and structured perturbations, a common setting across healthcare (e.g., drug response, disease progression, and other treatment-driven processes). The limitations of traditional pipelines, such as differential expression analysis, are ubiquitous in these settings: overly simplistic univariate and linear model assumptions, and a lack of discovery power by relying on incomplete knowledge bases. We demonstrate that causal discovery can be a complementary method that considers nonlinearity, multivariate interactions, and the distribution of the perturbation variable to identify driver variables and provide structure-based hypotheses of novel mechanisms.

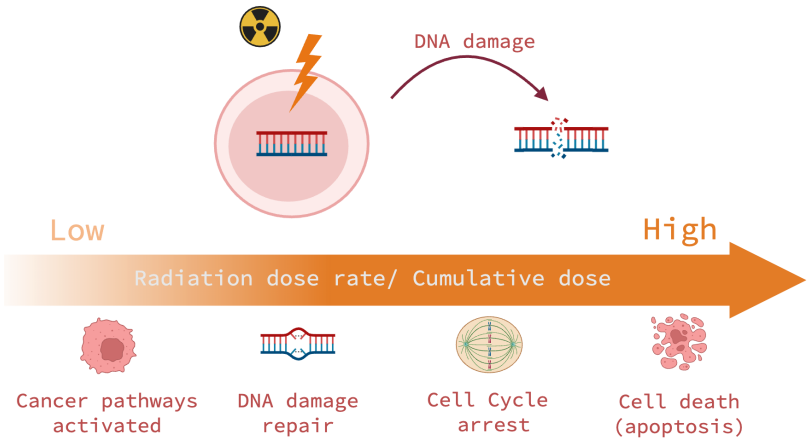

Figure 1: Exposure to ionizing radiation causes DNA breakage in cells. Cells activate pathways in order to repair the damage before the cell divides. When damage is excessive or occurs too rapidly, cells initiate programmed cell death (apoptosis).

2. Related Work

2.1. Finding Important Genes: Univariate and Multivariate Approaches

Differential expression analysis is the standard bioinformatics tool used to infer genes with expression levels that differ significantly between conditions (e.g., perturbed and control). This analysis tests the null hypothesis that the logarithmic fold change (LFC) between perturbed and control for a gene's expression is exactly zero, i.e., that the gene is not affected by the perturbation. The goal is to produce a list of genes passing multiple-test adjustment, ranked by p-value.

The most popular tool `DESeq2` (Love et al., 2014) employs this approach; each gene is modeled by a generalized linear model and negative binomial distribution. `DESEq2` models relationships between genes by assuming that genes of similar average expression have a similar dispersion; however, the statistical test for determining significant change between condition groups is univariate for each gene. Therefore, this method does not account for correlations and coordinated effects between groups of genes.

Multivariate tests including Hotelling's T^2 (Lu et al., 2005) and N-statistic (Klebanov et al., 2007) consider the joint distribution of expression levels. Glazko and Emmert-Streib (2009) evaluate univariate and multivariate tests and show that different tests find common but also complementing differentially expressed gene sets – each test projects on different aspects of the data. They propose use of both univariate and multivariate tests for the analysis of biological data for increased power.

Wenric and Shemirani (2018) propose multivariate classifiers to jointly model effects across genes and classify samples into perturbed and control groups. They find that random forest classifiers outperform differential expression analysis in 9 of the 12 cancer datasets evaluated. They further demonstrate that differential expression analysis might miss important genes, and a supervised learning-based gene selection method can be used to supplement these shortcomings.

2.2. Inferring Novel Pathways with Graph Learning

Once important genes are identified pathway enrichment is used to identify the processes that underlie response to perturbations (Raudvere et al., 2019). Pathway enrichment algorithms determine the significance of gene overlap to known pathways, recorded in knowledge bases like Gene Ontology (Consortium, 2019), KEGG (Kanehisa, 2002), Reactome (Milacic et al., 2024), and WikiPathways (Agrawal et al., 2024).

There is growing interest in developing algorithms that can identify novel pathways to guide the discovery of new mechanisms and biomarkers. Many gene regulatory network (GRN) inference algorithms have been proposed to solve this challenge. These methods include GENIE3 (Marbach et al., 2012) which uses an ensemble of random forest regression models to recover transcription factor (TF) to target gene directed relationships. GRNBoost2 (Moerman et al., 2019) takes a similar approach, but with gradient boosting which is much faster and more scalable. Methods like GENELink (Chen and Liu, 2022) and GRINDC (Feng et al., 2023) use deep learning to predict the presence or absence of an edge from embedded representations of single-cell RNA-sequencing datasets (Dixit et al., 2016). Several of these algorithms have been benchmarked against BEELINE (Pratapa et al., 2020). The authors show that GRN inference remains an open problem despite these recent advances.

Causal discovery methods infer directed acyclic graphs (DAGs) from data. These approaches benefit from theoretical guarantees of identifying the true causal graph under certain assumptions. A breakthrough in causal discovery was the definition of a continuous acyclicity constraint, allowing for gradient-based learning of DAGs (Zheng et al., 2018). This includes deep learning approaches such as DAG-GNN (Yu et al., 2019) which employs a variational autoencoder to learn an adjacency matrix under the DAG constraint while capturing nonlinear relationships between genes. A recent benchmark paper (Chevalley et al., 2025) evaluates several causal discovery algorithms with single-cell RNA-seq and Perturb-seq datasets, using knowledge graphs for edge validation. Their findings underscored issues with scalability and suboptimal use of perturbational data, and report low precision and recall of ground truth regulatory edges.

In this paper, we use causal discovery to learn causal graphs over the joint distribution of gene expression *and a perturbation variable*. Therefore, we learn pathways specific to perturbation response. Our goal is not to reconstruct all known regulatory relationships, which previous works have shown is notoriously difficult, but rather to propose an alternative approach for identifying important genes under perturbed conditions. A key advantage of our approach is the inference of a directed graph over variables, enabling the generation of hypotheses about novel response mechanisms. To our knowledge this is the first use of causal discovery in this setting.

3. Data Collection

The bulk RNA-seq dataset is described by Jantre et al. (2026). Here, we provide a brief overview of the procedure. RPE1 cells were grown in the lab and treated to low-dose ionizing radiation from a Cesium-137 gamma ray source at $D = 5$ different dose rates measured in

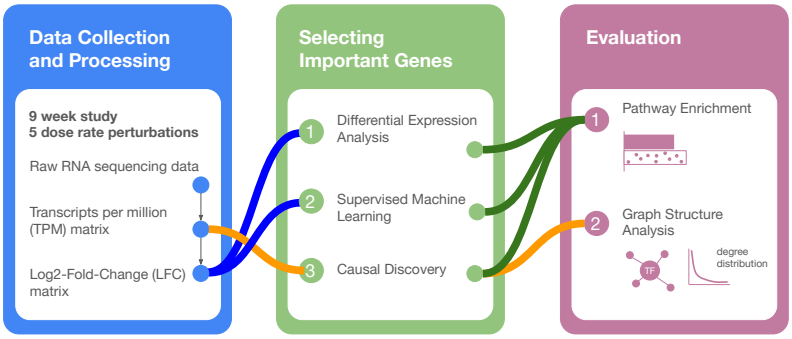

Figure 2: Our workflow comprises three main pipelines. Our contribution, shown in orange, is the use of causal discovery to identify important genes from transcriptomics data and downstream graph structural evaluation. We compare against standard pipelines for selecting important genes and subsequent pathway enrichment.

milliGray (mGy) per hour [1], including a control group with no radiation exposure. RNA-seq measurements were made for control and treated samples after each week for $T = 9$ weeks with 2 replicates per group for a total of $n = 2 \times (D + 1) \times T = 108$ samples. Raw RNA-seq reads were preprocessed with alignment to the human reference genome and filtered for low quality reads. Additionally, transcripts per million (TPM) values were computed per sample using `TPMCalculator v0.0.3`, resulting in a data matrix $X_{\text{TPM}} \in \mathbb{R}^{n \times p}$, where $p = 15,694$ genes. We denote $x_{(t,d),g}$ as a value in the matrix corresponding to week t, dose rate d (including $d = 0$ controls) and gene g. We compute the log2-Fold-Change matrix $X_{\text{LFC}} \in \mathbb{R}^{m \times p}$, where $m = T \times D = 45$ treated conditions. This matrix is computed element-wise for each gene : $X_{\text{LFC}_{(t,d),g}} = \log_2\left(\frac{x_{(t,d),g}}{x_{(t,0)g}}\right)$.

4. Methods

A visualization of our framework is shown in Fig. 2 which is comprised of three pipelines. In this section, we describe the second and third pipelines.

4.1. Selecting Important Genes

4.1.1. Differential Expression Analysis

Differential expression analysis was performed using `pyDESeq2 v0.4.9` with dataset X_{LFC} to determine which genes had statistically significant expression change in each of the 45 control-exposed pairs. Wald tests were performed and p-values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR). Genes with adjusted

1. One Gray is equal to one joule of absorbed radiation energy per kilogram of matter

p-value < 0.05 and $|X_{\text{LFC}_{(t,d),g}}| > 1$ were considered differentially expressed. For analysis, we aggregate differentially expressed genes (DEGs) at each dose rate $d \in D$ across time: $G_d = \bigcup_{t \in T} \{g : |X_{\text{LFC}_{(t,d),g}}| > 1, p\text{-value} < 0.05\}$.

4.1.2. Supervised Machine Learning

`RandomForestRegression` models were trained to predict the `(week,dose_rate)` labels of each sample in X_{LFC}. Weeks were assigned ordinal values $T = 1, 2, 3, 4, 5, 6, 7, 8, 9$ and dose rates were assigned $D = 0.28, 0.38, 0.55, 6.66, 12.11$ (mGy/hr) scalar values according to the measured exposure. We perform `RepeatedKFold` cross validation with 5 folds and 5 repeated runs. Within each fold, data is normalized to between (0,1]. The top 1000 features in each trained model's `feature_importance_` vector were extracted, and we selected the final set of genes that were stable in 20 out of 25 folds. Note that this gene set is dose agnostic since we use a stratified training split of `(week, dose_rate)` labels to train the models.

4.1.3. Causal Discovery

We use `DAG-GNN` (Yu et al., 2019), a variational autoencoder parameterized by a graph neural network, to construct DAGs at each dose rate from X_{TPM} and an additional column vector with the cumulative radiation of each sample, computed as `dose_rate` $\times$ `week` $\times$ 168 hours/week. We split X_{TPM} row-wise by dose rate so that X_{TPM_d} represents the samples collected at dose rate d. `DAG-GNN` training is time and memory intensive; to learn an adjacency matrix over $O(10,000)$ variables would be intractable. Therefore, we first filter out genes at each dose rate with p-value > 0.05 from the differential expression analysis, thereby removing genes that were not significantly expressed while still retaining genes with small LFC values. The resultant X_{TPM_d}'s at each dose rate after this preprocessing is shown in Table 1. We further partitioned the gene set according to a causal partition defined in Shah et al. (2025) using an initial undirected structure to obtain overlapping subsets of genes that retain consistency of causal discovery. For our work, we use the protein-protein interaction network in the STRING database as the initial undirected structure, and the maximum subset size in the partition of the structure was 2866 genes. A final DAG is obtained using the merge algorithm outlined in Shah et al. (2025) which is based on a union of the subgraphs. We estimate a DAG at each dose rate $d \in D$ corresponding to a subset of X_{TPM}. We bootstrap `DAG-GNN` over 10 resamples of the data with replacement to produce a distribution of DAGs at each dose rate. Further, we derive a consensus graph over the distribution of DAGs, where we take the union over edges that co-occur in 50% or more of the runs. To determine important genes at each dose rate, we take the nodes of the consensus graphs with at least one edge. In practice, many nodes end up being islands in the consensus graphs, which results in a gene set at each dose rate that is significantly smaller than the entire gene set $p = 15,694$. For a description of the partitioning algorithm, training procedure and selected hyperparameters see Appendix A.

Table 1: Dimensions of each X_{TPM_d} for each dose rate after pre-processing.

Dose Rate d (mGy/hr)	# Genes p	# Samples n
0.28	9262	36
0.38	9414	36
0.55	6274	36
6.66	8326	36
12.11	10603	36

4.2. Evaluation

4.2.1. Pathway Enrichment

After important genes are identified, pathway enrichment is used to understand the functional roles these genes play in the exposed condition. We use the Python interface for `gProfiler` for pathway enrichment analysis of genes identified from differential expression analysis, supervised machine learning with random forest regression models, and causal discovery. `gProfiler` performs functional profiling of gene lists using various sources of biological evidence (Gene Ontology terms, biological pathways, regulatory DNA elements, human disease gene annotations, and protein-protein interaction networks). `gProfiler` uses Fisher's one-tailed test as the p-value measuring the randomness of the intersection between the query gene set and the pathway term. The higher the $-\log_{10}p$-value the stronger the enrichment of that term is in the gene set. We provide a set of background genes (all genes in X_{TPM}) to ensure pathway enrichment p-values are comparable to each other across queries. We screen for pathways with `term_size` < 300 to filter out broad biological processes in the Gene Ontology that have thousands of genes.

4.2.2. Graph Structural Analysis

We validate the edge set and graph topologies of the DAG distributions inferred at each dose rate. We perform TF enrichment to determine if nodes with high out-degree (hub nodes) overlap with known TF genes which code for proteins that regulate hundreds of other genes. We use TRRUST (Han et al., 2018) and a ChIP-Seq network constructed with ENCODE (Davis et al., 2018) and Chip-Atlas (Zou et al., 2022) knowledge bases to identify human transcription factors and directed regulatory edges. Following the strategy of Chevalley et al. (2025) we use networks from a well studied cell line HepG2 because the RPE1 cell line has not been as comprehensively characterized (both cell lines are epithelial). We measure the overlap of edges in the DAGs with known protein-protein interactions (STRING Mering et al. (2003) and CORUM Giurgiu et al. (2019)), and regulatory edges (TRRUST, ChIP-seq network). We report the precision and F1-scores of the consensus graphs at each dose rate.

5. Results

Differential expression analysis with `DESeq2` and causal discovery with `DAG-GNN` both identify gene sets at each dose rate. Gene sets and intersections are visualized as Venn diagrams in

Fig. 3. Differential expression identifies larger sets in general, with the exception of dose rate 0.55 mGy/hr. Gene sets between methods have small overlap, the highest at dose rate 12.11 mGy/hr – these are also the largest gene sets for each method. Supervised machine learning with `RandomForestRegression` models yielded 68 genes inferred across all doses and have minimal overlap with causal graph and differential expression gene sets (Appendix Fig. 12).

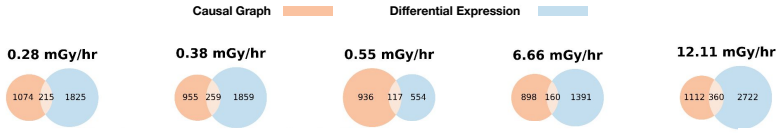

Figure 3: Venn diagrams showing gene overlaps between causal graphs and differential expression at each dose rate.

Pathway enrichment with `gProfiler` is shown in Fig. 4 for a set of manually curated radiation response specific pathways from the Gene Ontology, KEGG, WikiPathways and Reactome. Causal graph gene sets have higher enrichment across radiation response pathways, indicating that causal gene sets better identify important genes for radiation response with smaller gene sets. Supervised machine learning did not enrich any radiation pathways with the exception of `Genes and complexes involved in DNA repair pathways`. Certain pathways such as `Non-homologous end-joining` and `Cell cycle arrest` were not enriched in any method. We do not see a significant dose-dependent enrichment signal; that is, higher dose rates do not correspond to specific pathways compared to lower dose rates. We validate our enrichment results against random genes from the background set and genes that have high correlation with the cumulative radiation (Appendix Fig. 15).

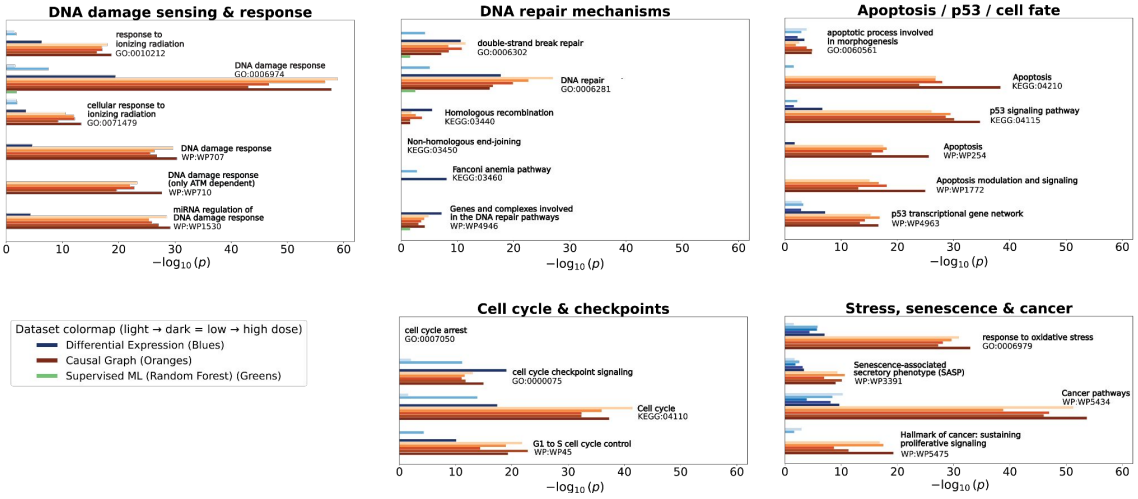

Figure 4: Pathway enrichment with `gProfiler` for manually curated radiation pathways across knowledge bases, categorized by radiation response type.

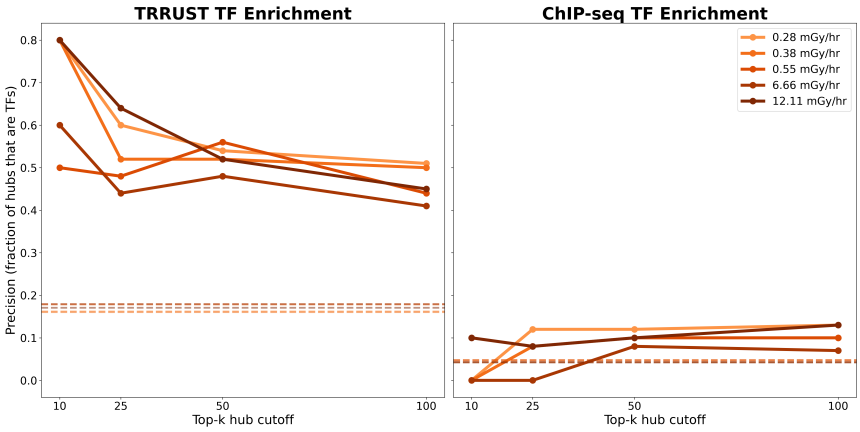

Figure 5: Percent of top-k out-degree nodes (hubs) that are known transcription factor genes for high out-degree nodes by dose rate for TRRUST and ChIP-seq knowledge bases. Dashed lines represent the baseline fraction of known transcription factors are in the graph's node set.

TF enrichment for each causal graph is shown in Figure 5. We see that for the top $k = 10$ hub genes in the causal graph, up to 80% are verified as genes that code for transcription factors in the TRRUST database. This percentage decreases as k increases; however, enrichment is always above the baseline random assignment of transcription factors to genes (shown in dashed lines). TF enrichment compared to the ChIP-Seq network shows significantly less overlap; this may be because of the mismatch in cell types since there is no comprehensive network for RPE1 cell lines.

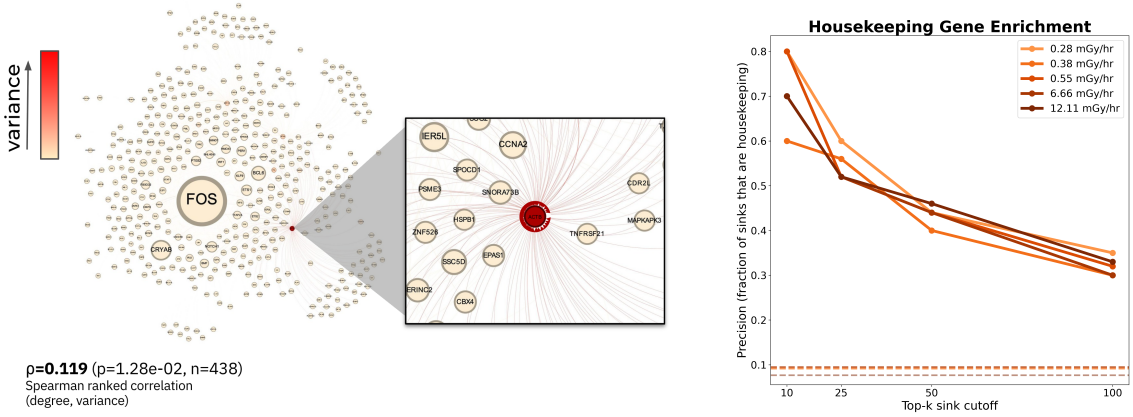

Figure 6: (**Left**) An example causal graph at dose rate 6.66 mGy/hr, with nodes colored by variance. Several highly connected sink nodes correspond to housekeeping genes; ACTB is shown in the zoomed-in panel. (**Right**) Percent of top-k in-degree nodes (sinks) that are housekeeping genes for each dose rate. Dashed lines represent the baseline fraction of housekeeping genes in the graph's gene set.

To rule out the possibility that causal discovery simply assigns highly variant genes as hub nodes, we compute the Spearman ranked correlation between node degree and variance in X_{TPM}; see Fig. 6 on the left. We find that the highest correlation from all dose rates is $\rho = 0.119$. This means there is a small tendency for higher-degree genes in the causal graph to have higher expression variance, but causal discovery is not just recapitulating high-variance genes and instead is finding biologically meaningful structure. This is evidenced by high out-degree nodes mapping to known TFs which are master regulators, and high in-degree nodes map to known housekeeping genes (Fig. 6 on the right) which are maintenance genes essential for cellular function. Housekeeping genes have also recently been shown to have high instability in stress conditions (Eisenberg and Levanon, 2013), which includes radiation exposure [2].

Table 2: Edge overlap between causal graphs and prior knowledge databases. TP = true positive number of edges found in the knowledge graph ; F1 = harmonic mean of precision and recall. For TRRUST and ChIP-seq, directed and reversed true positives are shown separately.

Dose		TRRUST			STRING		CORUM		ChIP-seq		
(mGy/hr)	Edges	TP_{dir}	TP_{rev}	F1	TP	F1	TP	F1	TP_{dir}	TP_{rev}	F1
0.28	2777	24	12	0.012	143	0.026	54	0.020	111	107	0.006
0.38	2610	17	7	0.009	130	0.025	47	0.019	112	80	0.006
0.55	1248	3	4	0.003	52	0.016	26	0.019	54	16	0.004
6.66	2263	16	8	0.010	113	0.031	34	0.019	88	82	0.006
12.11	2417	18	13	0.009	142	0.025	36	0.015	107	89	0.004

In Table 2, we show the amount of edge overlap between causal graph at each dose rate and different knowledge graphs, including undirected protein-protein interaction graphs or complexes (STRING, CORUM) and directed regulatory graphs (TRRUST, ChIP-Seq). We see minimal overlap, with a maximum true positive edge count of 143, and maximum F1 score of 0.031. This is similar to results from Chevalley et al. (2025), which used larger scale single-cell datasets for causal discovery.

The set of "invariant" causal graph nodes is the intersection of nodes across all dose rates and results in 438 genes. We performed pathway enrichment on this invariant gene set, which we hypothesized as encoding fundamental radiation response relationships. The top 10 enriched pathways are shown in Table 3. Note that unlike previous enrichment analyses, here we report the top pathways without explicit manual curation of radiation-specific pathways. Despite this, all top pathways correspond to well-known radiation response pathways (Sasaki et al., 2002; Feinendegen et al., 2004). We include the top 10 pathways for differential expression which have some overlap to radiation response (specifically inflammation pathways and tumor necrosis) but have larger p-values in Appendix Table 5.

2. Housekeeping gene set from Hsiao et al. (2001).

Table 3: Top 10 Enriched Pathways for the Causal Invariant Gene Set ($n = 438$)

Pathway	$-\log_{10}(p)$	Term Size	Intersection Size
Cell cycle WP:WP179	26.76	120	31
p53 signaling pathway KEGG:04115	26.48	74	26
Cell cycle KEGG:04110	25.94	157	33
miRNA regulation of DNA damage response WP:WP1530	23.34	73	24
signal transduction by p53 class mediator GO:0072331	23.02	164	31
Interleukin-4 and Interleukin-13 signaling REAC:R-HSA-6785807	22.70	111	27
DNA damage response WP:WP707	22.54	69	23
Integrated breast cancer pathway WP:WP1984	22.45	152	30
Cellular senescence KEGG:04218	22.43	155	30
Mitotic G1 phase and G1/S transition REAC:R-HSA-453279	21.65	148	29

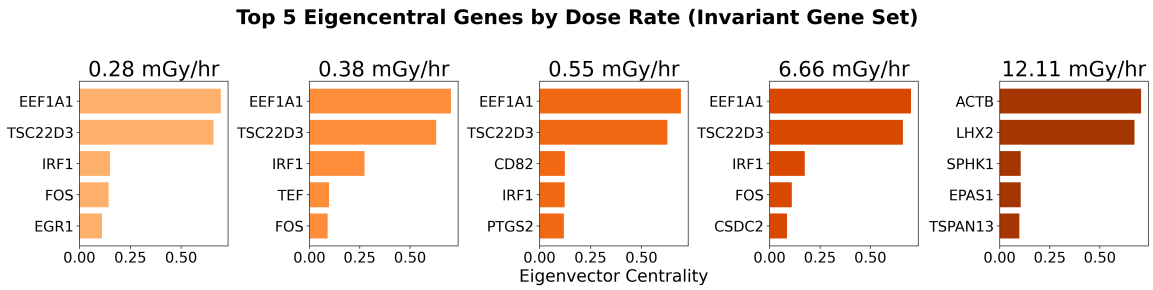

Figure 7: The top 5 identified eigencentric genes for each dose-rate causal subgraph induced by the invariant set of genes.

We identify causal subgraphs over the invariant gene set at each dose rate and compute the top eigencentral genes (a measure of the node's influence in the subgraph) in Fig. 7. We find that top scoring genes correspond to transcription factors (FOS, TSC22D3, IRF1, EGR1, TEF, LHX2, EPAS1) and housekeeping genes (EEF1A1, ACTB), again validating the known structure regarding these genes. Notably, the highest dose-rate (12.11 mGy/hr) has different eigencentric genes suggesting a different biological structure compared to the lower dose rates. We additionally visualize a subset of edges in the subgraphs over the invariant gene set at each dose rate in Fig. 8 corresponding to "perfect" edges that co-occur across 100% of the bootstrap distribution. These correspond to stable edges identified by `DAG-GNN` and, although only a handful are true positives in knowledge bases, these edges could be candidates for novel relationships with biological significance in radiation response.

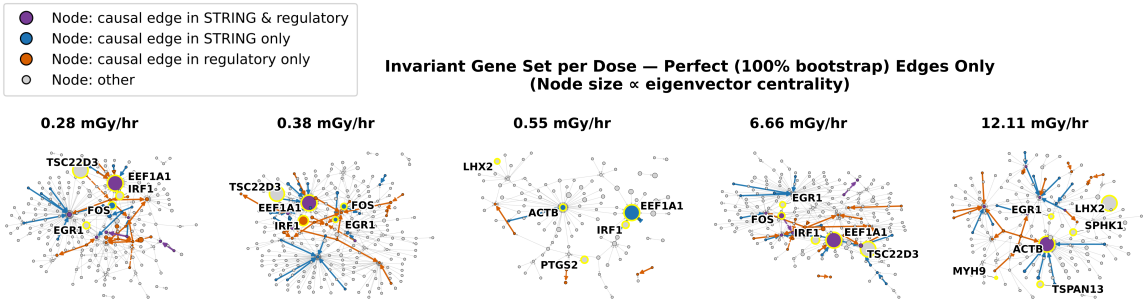

Figure 8: The invariant subgraphs at each dose rate with only the "perfect" edges that co-occur across 100% of the bootstrap distribution visualized. Annotated genes correspond to the top-5 eigencentric genes at each dose rate. Edges are highlighted according to true positives in knowledge bases where "regulatory" means either TRRUST or ChIP-seq edges.

Housekeeping genes had a high overlap in the causal invariant gene set (14.4%) compared to the differential expression invariant gene set (1.5%). In Fig. 9, we perform pathway enrichment on the ancestors compared to non-ancestors of housekeeping genes. We find that the enrichment profile between these two groups is distinct: the ancestor set overlaps with response to oxidative stress (ROS) pathways, and the non-ancestor set overlaps with membrane permeability pathways. This suggests a branching effect between stress (ROS) and apoptosis (membrane permeability is an indicator of triggered cell death). This branching effect is well documented in the literature (Fulda et al., 2010).

The results reported in this section use the consensus causal graphs over `DAG-GNN` bootstrap runs, which filter for edges that co-occur in more than 50% of the bootstrap runs. Analyses for each DAG in the bootstrap is in Appendix G.

6. Discussion

We present a previously unexplored application of causal discovery with transcriptomics data to identify important genes related to perturbation response. Typically differential expression analysis is performed to identify important genes by comparing a control group to perturbed groups. These genes are then used for pathway enrichment to identify possible biomarkers of the perturbation condition. We demonstrate that differential expression

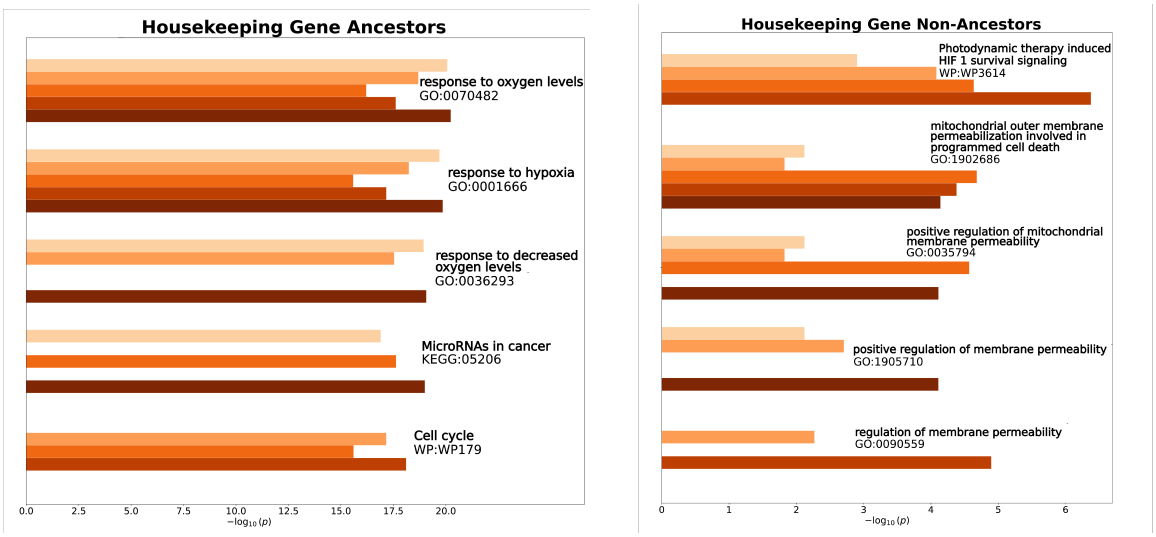

Figure 9: (**Left**) Enrichment of housekeeping genes in the parent sets of the causal subgraphs over the invariant gene set. (**Right**) Enrichment of housekeeping genes among the non-ancestors in the causal subgraphs over the invariant gene set.

analysis may potentially miss important genes related to perturbation response. By jointly modeling the perturbation variable and the multivariate gene expression distribution we show that causal discovery has a higher hit rate of finding perturbation response related gene sets, using radiation response as a use case. Furthermore, the learned DAG identifies structural signatures such as transcription factors and housekeeping genes and hub and sink nodes respectively. We find that causal subgraphs induced by a core set of invariant genes are enriched for perturbation-specific pathways and appear to encode branching between apoptosis and stress response pathways. This observation aligns with invariant causal prediction (Peters et al., 2016), which proposes that causal relationships exhibit stability across environments, here corresponding to different dose rates. To infer novel pathways with causal discovery, one can identify edges that co-occur across the entire bootstrap distribution corresponding to stable edges with potential biological significance. These novel pathways should then be validated with experimentation, for example, with CRISPR knockouts of upstream genes.

Applications of causal discovery have focused on exact recovery of causal edges; this is a notoriously difficult task when we consider the numerous ways in which most realistic datasets violate assumptions needed for identifiability. Research currently focuses on collecting large-scale perturbational datasets to improve identification of the true causal graph. However, there is a need for interpretable methods for biomarker and mechanism discovery in low-sample settings (e.g., small patient cohorts, rare diseases, or settings in which high throughput data is difficult to collect). In the presence of strong upstream perturbations (e.g., environmental stress or drug dosage), we propose that causal discovery should be considered as part of the bioinformatics toolbox as existing methods may be suboptimal in this setting.

Limitations Our focus on radiation response may limit the generalizability of our findings. Future work will extend this analysis to additional transcriptomic datasets to evaluate the broader applicability of causal discovery for modeling perturbation response. We note that causal discovery incurs substantially higher computational cost than differential expression analysis and supervised machine learning (Appendix A); this limits large-scale statistical evaluation. As a result, our current analysis is primarily qualitative, relying on consensus graphs across bootstrap runs. A more rigorous quantification of these findings through statistical analysis over distributions of consensus graphs is an important direction for future work. Finally, we acknowledge that pathway enrichment is biased toward well-studied genes and can be prone to over-interpretation.

Acknowledgments

We thank Becca Weinburg for helpful discussions and explanations of experimental protocols. AS is supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. This research used resources of the Argonne Leadership Computing Facility. Argonne National Laboratory's work on the Exploration of the Potential for Artificial Intelligence and Machine Learning to Advance Low-Dose Radiation Biology Research (RadBio-AI) project was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environment Research, under contract DE-AC02-06CH11357.

References

Ayushi Agrawal, Hasan Balcı, Kristina Hanspers, Susan L Coort, Marvin Martens, Denise N Slenter, Friederike Ehrhart, Daniela Digles, Andra Waagmeester, Isabel Wassink, et al. Wikipathways 2024: next generation pathway database. *Nucleic acids research*, 52(D1): D679–D689, 2024.

Philippe Brouillard, Chandler Squires, Jonas Wahl, Konrad P Kording, Karen Sachs, Alexandre Drouin, and Dhanya Sridhar. The landscape of causal discovery data: Grounding causal discovery in real-world applications. *arXiv preprint arXiv:2412.01953*, 2024.

Guangyi Chen and Zhi-Ping Liu. Graph attention network for link prediction of gene regulations from single-cell rna-sequencing data. *Bioinformatics*, 38(19):4522–4529, 2022.

Mathieu Chevalley, Yusuf H Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. A large-scale benchmark for network inference from single-cell perturbation data. *Communications Biology*, 8(1):412, 2025.

Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.

Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, et al. The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research*, 46(D1):D794–D801, 2018.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.

Eli Eisenberg and Erez Y Levanon. Human housekeeping genes, revisited. *TRENDS in Genetics*, 29(10):569–574, 2013.

Ludwig E Feinendegen, Myron Pollycove, and Charles A Sondhaus. Responses to low doses of ionizing radiation in biological systems. *Nonlinearity in biology, toxicology, medicine*, 2(3):15401420490507431, 2004.

Ke Feng, Hongyang Jiang, Chaoyi Yin, and Huiyan Sun. Gene regulatory network inference based on causal discovery integrating with graph neural network. *Quantitative Biology*, 11(4):434–450, 2023.

Simone Fulda, Adrienne M Gorman, Osamu Hori, and Afshin Samali. Cellular stress responses: cell survival and cell death. *International journal of cell biology*, 2010(1):214074, 2010.

Madalina Giurgiu, Julian Reinhard, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. Corum: the comprehensive resource of mammalian protein complexes—2019. *Nucleic acids research*, 47(D1): D559–D563, 2019.

Galina V Glazko and Frank Emmert-Streib. Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics*, 25(18):2348–2354, 2009.

Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, et al. Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, 46(D1):D380–D386, 2018.

Li-Li Hsiao, Fernando Dangond, Takumi Yoshida, Robert Hong, Roderick V Jensen, Jatin Misra, William Dillon, Kailin F Lee, Kathryn E Clark, Peter Haverty, et al. A compendium of gene expression in normal human tissues. *Physiological genomics*, 7(2):97–104, 2001.

Sanket Jantre, Kriti Chopra, Guang Zhao, Clark Cucinell, Rebecca Weinberg, Sara Forrester, Thomas Brettin, Nathan M Urban, Xiaoning Qian, and Byung-Jun Yoon. Interpretable transcriptome-to-phenotype modeling of cell-painting nuclear morphology features from rna-seq under low-dose radiation exposure. *bioRxiv*, pages 2026–02, 2026.

Minoru Kanehisa. The kegg database. In *'In silico'simulation of biological processes: Novartis Foundation Symposium 247*, volume 247, pages 91–103. Wiley Online Library, 2002.

Lev Klebanov, Galina Glazko, Peter Salzman, Andrei Yakovlev, and Yuanhui Xiao. A multivariate extension of the gene set enrichment analysis. *Journal of bioinformatics and computational biology*, 5(05):1139–1153, 2007.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

Yan Lu, Peng-Yuan Liu, Peng Xiao, and Hong-Wen Deng. Hotelling's t 2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, 21(14):3105–3113, 2005.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, et al. The reactome pathway knowledgebase 2024. *Nucleic acids research*, 52(D1):D672–D678, 2024.

Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, 2019.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and andT M Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.

Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198, 2019.

Diletta Rosati, Maria Palmieri, Giulia Brunelli, Andrea Morrione, Francesco Iannelli, Elisa Frullanti, and Antonio Giordano. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and structural biotechnology journal*, 23:1154–1168, 2024.

Masao S Sasaki, Yosuke Ejima, Akira Tachibana, Toshiko Yamada, Kanji Ishizaki, Takashi Shimizu, and Taisei Nomura. Dna damage response pathway in radioadaptive response. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 504(1-2): 101–118, 2002.

Ashka Shah, Adela Frances DePavia, Nathaniel C Hudson, Ian Foster, and Rick Stevens. Causal discovery over high-dimensional structured hypothesis spaces with causal graph partitioning. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=FecsgPCOHk.

Yukiko Shimizu, Kazunori Kodama, Nobuo Nishi, Fumiyoshi Kasagi, Akihiko Suyama, Midori Soda, Eric J Grant, Hiromi Sugiyama, Ritsu Sakata, Hiroko Moriwaki, et al. Radiation exposure and circulatory disease risk: Hiroshima and nagasaki atomic bomb survivor data, 1950-2003. *Bmj*, 340, 2010.

Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

Marcel Verheij and Harry Bartelink. Radiation-induced apoptosis. *Cell and tissue research*, 301(1):133–142, 2000.

Stephane Wenric and Ruhollah Shemirani. Using supervised learning methods for gene selection in rna-seq case-control studies. *Frontiers in genetics*, 9:364621, 2018.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International conference on machine learning*, pages 7154–7163. PMLR, 2019.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Zhaonan Zou, Tazro Ohta, Fumihito Miura, and Shinya Oki. Chip-atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating chip-seq, atac-seq and bisulfite-seq data. *Nucleic acids research*, 50(W1):W175–W182, 2022.

Appendix A. `DAG-GNN` training at Scale

`DAG-GNN` is trained using a nested optimization loop where an inner loop updates the neural network parameters and adjacency matrix to minimize the evidence lower bound (ELBO) loss, while an outer loop enforces the DAG constraint using an augmented Lagrangian. A trainable adjacency matrix is represented by $A \in \mathbb{R}^{p\times p}$. The encoder is $z = (I - A^T)x$ for input $x \in \mathbb{R}^{n\times p}$, where I is the identity matrix. The decoder is $(I - A^T)^{-1}z$. We set the embedding dimension to 64. We trained `DAG-GNN` for 300 epochs using Adam with a learning rate of 3e-3 (`gamma`=1.0, `lr_decay`= 200). We use n for the batch size, because we do not have sufficient samples for stable mini-batching. The maximum number of iterations for the Lagrangian optimization was set to 100. The threshold for the adjacency matrix A is set to 0.3, which is the default value.

`DAG-GNN` training is time and memory intensive. To learn an adjacency matrix over $p = 15,694$ would be intractable. We partitioned the gene set according to a causal partition (Shah et al., 2025) of the protein-protein interaction network from the STRING database, then merged the graphs by taking the union of edges. Training times depend on the gene subset size, which vary by dose rate. The average training time for `DAG-GNN` on one subset of data across bootstraps was 23.3 min. Training curves for one example subset are shown in Fig. **??**. Note that loss spikes occur when the augmented Lagrangian penalty ρ is increased in the outer loop, abruptly strengthening the acyclicity constraint and perturbing the learned graph. This is expected behavior as the model rebalances reconstruction accuracy with enforcing a valid DAG structure. Models were trained on a NVIDIA Tesla V100-SXM2-32GB GPU.

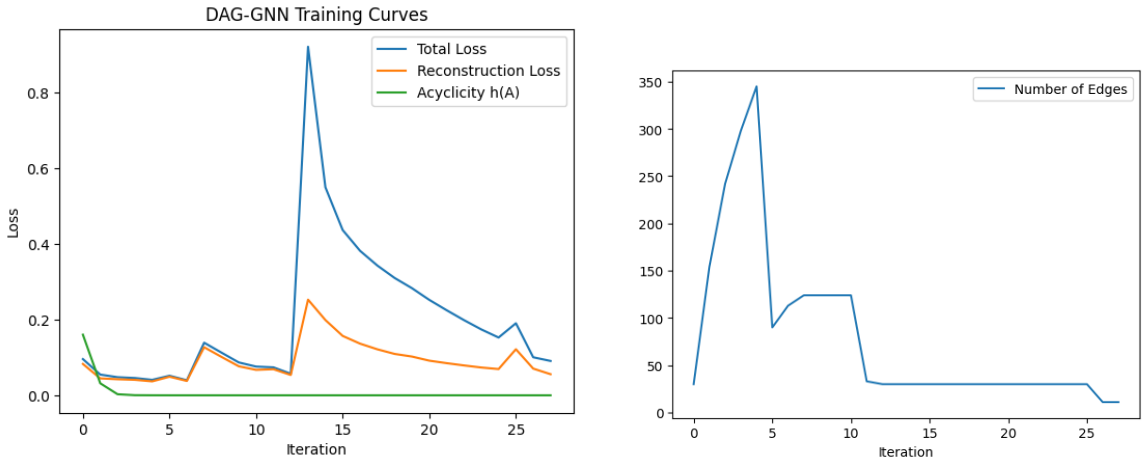

Figure 10: (**Left**) Loss curves and acyclicity constraint $h(A)$ which converges to 0 to ensure a DAG. (**Right**) The number of edges in the learned adjacency matrix at each iteration using the default threshold 0.3.

Appendix B. Linear Regression Baseline

We were concerned that causal gene sets might just be identifying genes that have strong correlative structure with the radiation vector, especially given a small sample size. We constructed a baseline algorithm based on linear regression of (week, dose_rate) labels onto each gene. Each model's R^2 measures how much variance in that gene's expression is explained by dose rate and week jointly. A gene with a high multiple correlation R^2 is one whose expression varies strongly as a linear function of dose rate and/or time. This captures genes that respond to radiation in a dose- or time-dependent manner — but purely through univariate linear association with the experimental conditions, without any causal graph structure or model-based feature importance. It serves as a "correlative baseline" against which the causal graph gene selections are compared. Similar to the supervised machine learning framework, we choose the 1000 genes at each fold with the top R^2 scores. Then we select genes that are stable in more than 50% of folds. This results in 481 genes and is visualized as a "Correlation" baseline in Fig. 11. We also perform pathway enrichment of this gene set and report the top 10 pathways in Table 4. We see enrichment of radiation pathways as well, but a much higher enrichment for the invariant causal gene set (Table 3 which also has a smaller gene set size.

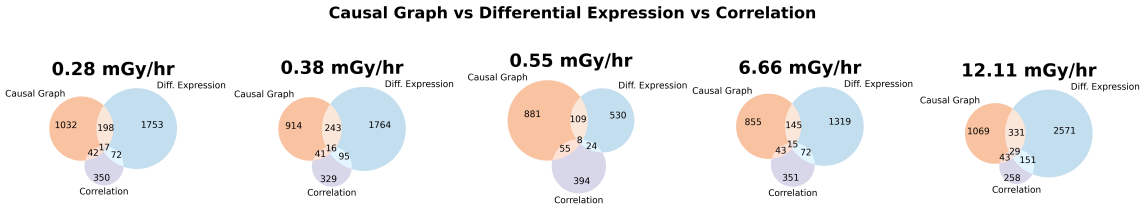

Figure 11: Intersections with the Correlation baseline algorithm (based on Linear Regression).

Appendix C. Additional Gene Set Intersections

There were 68 stable genes inferred by the `RandomForestRegression` models across folds. Intersections with causal gene sets and differential expression are shown in Fig. 12. Fig. 13 and Fig. 14 show within causal graph and differential expression intersections respectively. There are a core set of "invariant" genes across dose rates for each method which may correspond to genes involved fundamental radiation response processes. The invariant gene set is larger for causal graphs than differential expression (438 genes compared to 329 genes).

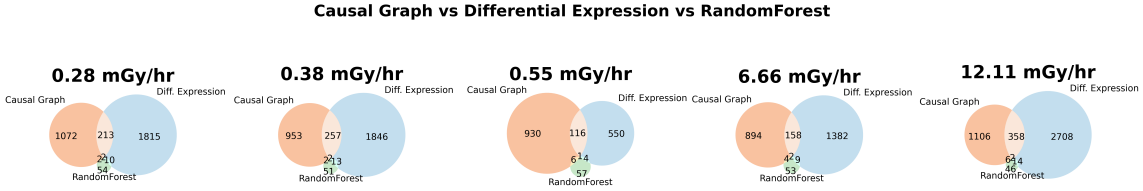

Figure 12: Intersections with the Random Forest gene set which is dose invariant.

Table 4: Top 10 Enriched Pathways — Correlation Gene Set

Pathway	$-\log_{10}(p)$	Term Size	Intersection
Cell Cycle Checkpoints REAC:R-HSA-69620	8.28	289	24
Regulation of TP53 Activity through Phosphorylation REAC:R-HSA-6804756	7.70	92	14
Regulation of TP53 Activity REAC:R-HSA-5633007	7.18	160	17
G1 to S cell cycle control WP:WP45	7.15	64	12
Cell cycle WP:WP179	6.96	120	15
Mitotic G1 phase and G1/S transition REAC:R-HSA-453279	6.82	148	16
Retinoblastoma gene in cancer WP:WP2446	6.79	89	13
Cell cycle KEGG:04110	6.30	157	16
DNA damage response WP:WP707	6.02	69	11
DNA replication GO:0006260	5.84	279	20

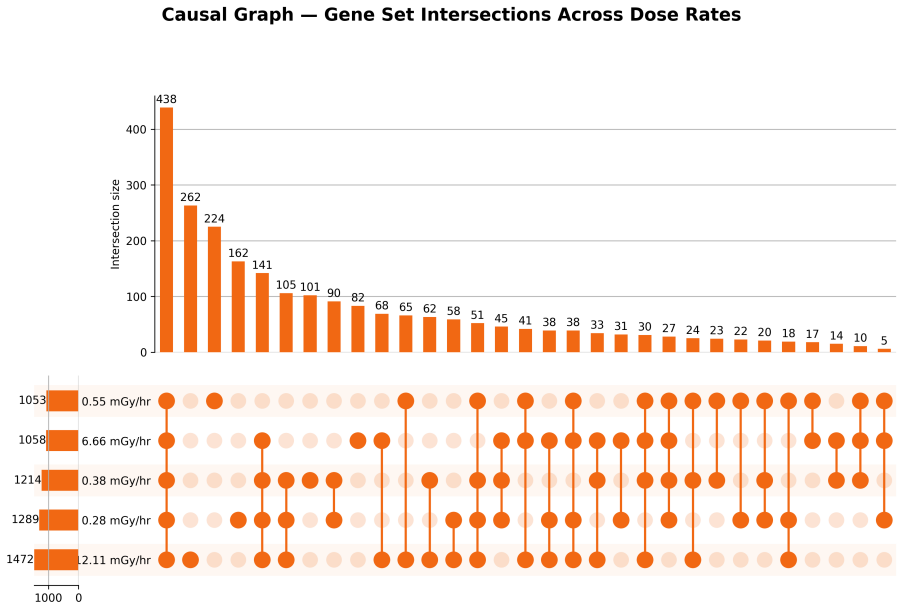

Figure 13: Intersections of causal graph gene sets across each dose rate. There is an "invariant" set of 438 genes which is the intersection across all dose rates.

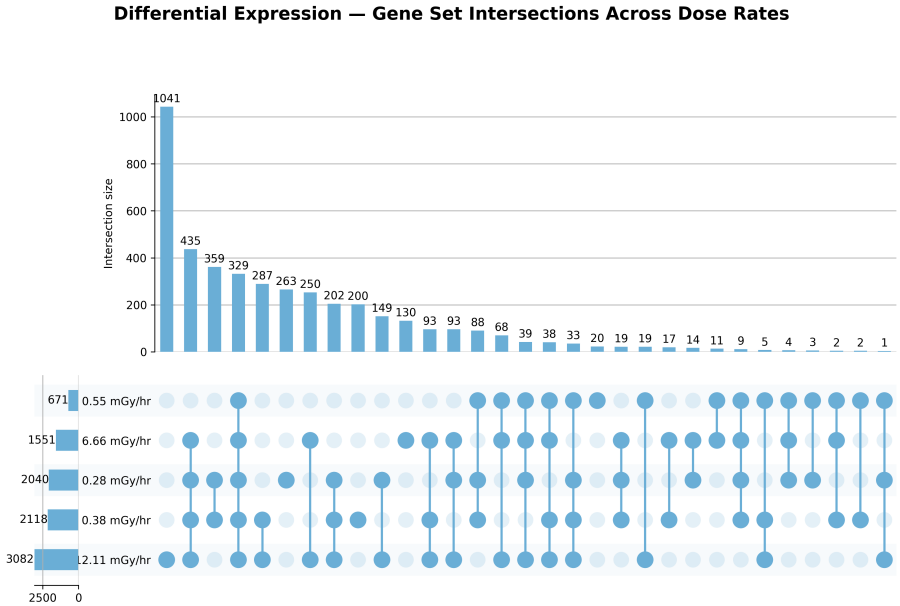

Figure 14: Intersections of differential expression graph sets across each dose rate. There is an "invariant" set of 329 genes which is the intersection across all dose rates.

Appendix D. Random and Correlative Gene Sets

Pathway enrichment $-\log_{10}p$-values were surprisingly large for causal gene sets (Fig. 4). To check that these values were meaningful, we ran a simple test with `gProfiler` with random genes taken from the background gene set (all genes measured in the experiment). In Fig. 15 we report the distribution of the top 10 enriched pathways (with 5 repeated runs). We verify random gene sets do not enrich any pathways. We also check our correlative linear model and see that these gene sets do enrich pathways, but not at the level of the causal gene sets. We already demonstrate in Fig. 4 that the causal gene sets perform better compared differential expression and `RandomForest` gene sets.

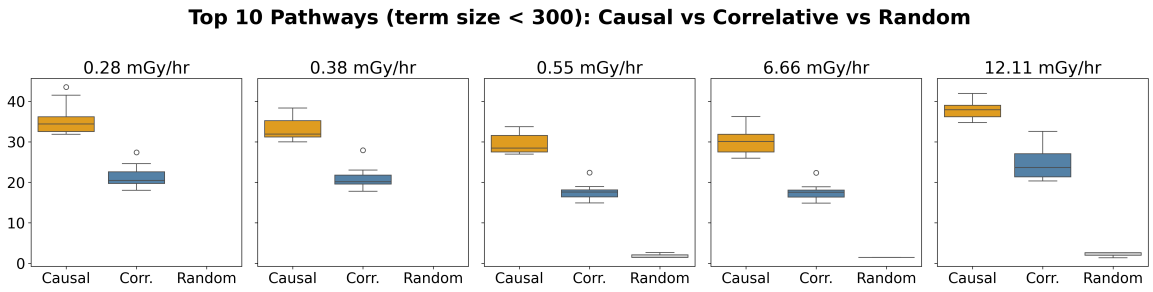

Figure 15: A sanity check for pathway enrichment checked enrichment of the top 10 pathways from the causal gene sets, correlative gene sets and random gene sets.

Appendix E. Top 10 Pathways for Differential Expression Invariant Gene sets

We report the top 10 pathways for the differential expression invariant gene set (329 genes) in Table 5.

Appendix F. Cluster Functional Analysis

We performed a cluster analysis of the causal graphs at each dose rate; an example clustering is shown in Fig. 16. To understand if each cluster is biologically meaningful and encodes distinct processes, we again perform pathway enrichment on each gene subset. We also compute overlaps to CORUM protein complexes which are manually annotated, experimentally validated, high quality annotations. In Table 6 we see significant enrichment of pathways and CORUM complexes in each cluster for dose rate 6.66 mGy/hr. Only C2 (Eukaryotic translation) and C3 (ATP- DNA binding), however, show clear consensus between the top pathway and the CORUM complex.

Table 5: Top 10 Enriched Pathways — Differential Expression Invariant Gene Set

Pathway	$-\log_{10}(p)$	Term Size	Intersection
cellular response to interleukin-1 GO:0071347	9.73	84	13
response to interleukin-1 GO:0070555	9.40	110	14
cellular response to tumor necrosis factor GO:0071356	7.68	208	16
glycosaminoglycan binding GO:0005539	7.15	249	17
response to tumor necrosis factor GO:0034612	7.09	229	16
glomerulus development GO:0032835	6.97	71	10
active monoatomic ion transmembrane transporter activity GO:0022853	6.71	117	12
Diabetic cardiomyopathy KEGG:05415	6.45	202	15
regulation of angiogenesis GO:0045765	6.42	293	17
nephron development GO:0072006	6.40	160	13

Table 6: Cluster enrichment for dose rate I (6.66 mGy/min). CORUM complex overlap identified by Fisher's exact test; top pathway from gProfiler enrichment analysis.

Cluster	CORUM Complex	p-value	Top Pathway	p-value
C0	DNA binding; chromosome; DNA topological change	4.81×10^{-5}	Antigen Presentation: Folding, assembly and peptide loading of class I MHC	8.10×10^{-21}
C1	Cell cycle checkpoint signaling; DNA binding; nucleus	1.26×10^{-4}	Pathogenic *Escherichia coli* infection	5.09×10^{-7}
C2	Cytoplasm; translation	2.94×10^{-24}	Eukaryotic Translation Elongation	6.66×10^{-22}
C3	ATP binding; chromosome; DNA replication	7.66×10^{-10}	ATP-dependent activity, acting on DNA	8.70×10^{-9}
C4	mRNA splicing, via spliceosome; spliceosomal complex	1.93×10^{-5}	Mitotic G1 phase and G1/S transition	1.79×10^{-16}
C5	Protein targeting; intracellular protein transport; endocytosis	1.82×10^{-2}	NRF2 pathway	3.91×10^{-6}
C6	Myeloid cell differentiation; regulation of apoptotic process	9.58×10^{-9}	Integrated breast cancer pathway	9.83×10^{-10}

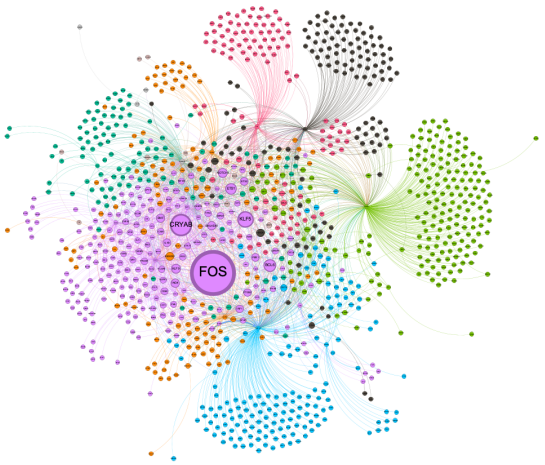

Figure 16: Example of the causal graph at dose rate 6.66 mGy/hr colored by cluster ID.

Appendix G. Causal Discovery Bootstrap Results

Here, we show the `DAG-GNN` metrics across bootstrap runs. The results in the main paper report metrics for consensus graphs formed by filtering for edges that co-occur in 50% or more of the bootstrap runs. We ran `DAG-GNN` with 10 bootstraps and metrics are shown in Fig. 17. Compared to Table 2 and Fig. 5, we see a high variance distribution and overall poorer results which supports our choice of choosing consensus edges for downstream analysis. In Fig. 18, we visualize the distribution of edge frequencies for subgraphs induced by the invariant causal gene set. We see that this subgraph has a higher density of "perfect" edges that appear in all bootstraps. This shows that the invariant gene sets and subgraph edges have a higher confidence, and this is supported by enrichment of only radiation pathways. These subgraphs and high frequency edges can be used as a starting point for hypothesized novel pathways, however, this requires experiments to prove.

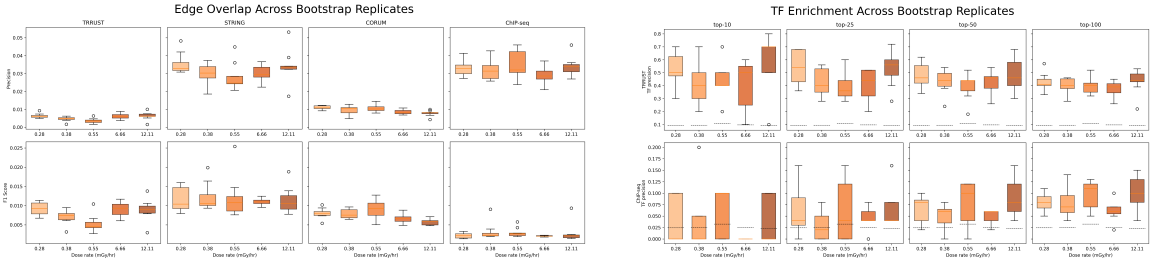

Figure 17: (**Left**) The distribution of edge overlap with knowledge bases for all graphs across 10 bootstrap runs. (**Right**) The percentage of top-k hub nodes that correspond to known transcription factors for all graphs in the 10 bootstrap runs.

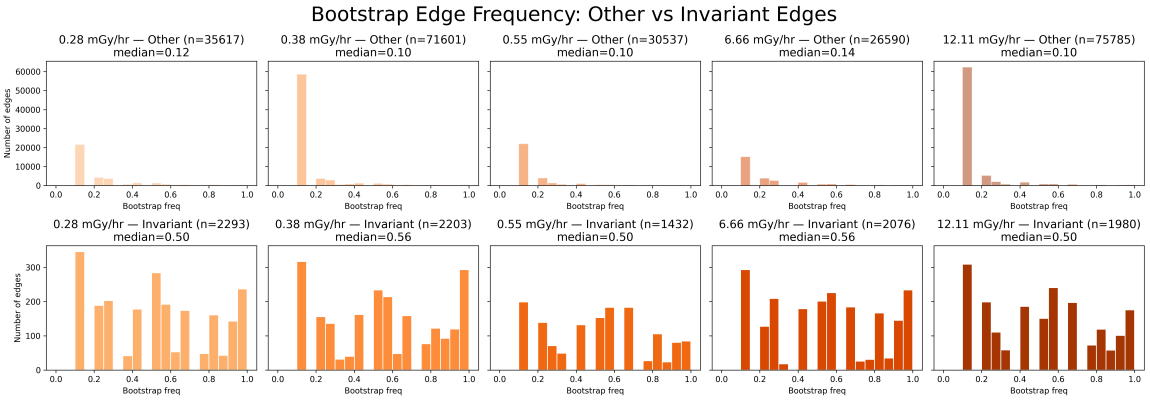

Figure 18: (**Top**) The distribution of edge frequencies at each dose rate across bootstraps. A frequency of 1.0 means that the edge appeared in all 10 of the DAGs. (**Bottom**) The distribution of edge frequencies at each dose rate subgraph induced by the causal invariant gene set (438 genes).