Elucidating Complex Biological Interactions Using Computational Techniques

by

Zhenjiang Fan

Bachelor of Science, Tianjin University of Technology, 2011

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Zhenjiang Fan

It was defended on

July 31, 2023

and approved by

Hyun Jung Park, Committee Co-Chair, Assistant Professor, Department of Human Genetics

Stephen Lee, Committee Co-Chair, Assistant Professor, Department of Computer Science

Heng Huang, Committee Member, John A. Jurenko Endowed Professor, Department of Electrical and Computer Engineering

Adriana Kovashka, Committee Member, Associate Professor, Department of Computer Science

Xulong Tang, Committee Member, Assistant Professor, Department of Computer Science

Elucidating Complex Biological Interactions Using Computational Techniques

Zhenjiang Fan, PhD

University of Pittsburgh, 2023

Studying complex biological systems faces numerous technical challenges due to their intricate nature and the multitude of interacting factors involved while analyzing related datasets. These challenges include the data diversity in biomedical datasets, nonlinearity behaviors among variable interactions, contextual causal factors, subtype heterogeneity, and causal mechanism complexity. To address these challenges, we must build specified computational models to tackle certain problems. Two of the most widely used computational tools are machine learning (ML) and deep learning (DL). Despite their tremendous potential, integrating ML and ML into biological research is not a trivial task.

In our first project where we aim to understand dynamics among complex biological networks, such as subtype biological networks for a disease, we utilized a network similarity measuring method based on normalized Laplacian matrix eigenvalue distribution to systematically identify a comparable estrogen receptor negative (ER-) normal ceRNA network comparable to estrogen receptor positive (ER+) normal reference ceRNA network. We exploited various network analysis techniques to study dynamics among constructed subtypes of breast cancer. Our systematically analyzing disease subtype network using these network analysis techniques provides a meaningful research direction.

For our second project where we determine to address the nonlinearity behavior and identify complex causal mechanisms in complex biomedical data, we developed a causal inference

method that learns both linear and nonlinear causal relations and estimates the effect size using a deep-neural network approach coupled with the knockoff framework. By using both simulation data and multiple real world biomedical datasets, we demonstrated that our proposed method outperforms existing methods in identifying true and known causal relations. The identified nonlinear causal relations and estimating their effect size can help understand the complex disease pathobiology, which is not possible using other methods.

In our third project where we aim to address the data diversity, nonlinearity behavior, contextual causal factor problems in single-cell sequencing datasets, we created a DL model to identify condition-specific cell subtypes when we have multiple types of information. In comparison with existing clustering algorithms, our proposed clustering method outperforms them in terms of various evaluation matrices using both simulation data and real-world single-cell sequencing data.

Table of Contents

List of Tables

List of Figures

Preface

With great pride as I am about to present this doctoral dissertation, I would like to express my gratitude to everyone who has helped me along this journey. It is a privilege and an honor to have been granted the opportunity to delve deep into a subject of personal passion to unravel its complexities. This dissertation would not have been possible without the guidance, support, and encouragement of numerous individuals who have played a crucial role in my academic journey.

First and foremost, I would like to express my deepest gratitude to my advisors, Professor Hyun Jung (HJ) Park and Professor Stephen Lee, whose expertise, wisdom, and unwavering commitment have been instrumental in shaping the trajectory of my research program. Their guidance has not only challenged me to think critically but also inspired me to push the boundaries of my intellectual capabilities.

Next, I would like to thank the members of my dissertation committee, John A. Jurenko Endowed Professor Heng Huang, Professor Adriana Kovashka, and Professor Xulong Tang, whose invaluable insights and constructive feedback have enriched my research program. Their collective expertise and rigorous examination have undoubtedly enhanced the rigor and quality of this dissertation.

Furthermore, I must acknowledge the Computer Science department and the University of Pittsburgh that have provided the necessary resources and opportunities for me to pursue my research. Additionally, I would like to thank Professor Taieb Znati as he helped me when I desperately needed some support and advice; I would like to thank Professor Qi Mi and Professor Joseph Carcillo, as each of them helped me with one of my research projects; I would like to thank all the professors and staff members, like Keena Walker and others, here at the Computer Science

department for their help; and I would also like to thank my peers at the department (like Nannan Wen, Debarun Das, Kevin Hostler, Nathan Ong, and others), my lab peers (Yulong Bai, Yidi Qin, and others), and my friends.

Lastly, I am grateful to my wife (Amber York), my parents (Aiyun Gao and Jisheng Fan), my sister (Lingling Fan), my brother (Zhendong Fan) and his family (my sister-in-law, my nieces, and nephew), and my wife's family (Bert York, Irene York, and Brian York), for their unwavering belief in me and their unwavering support throughout this arduous endeavor. Their love, encouragement, and understanding have sustained me during moments of doubt and fueled my determination to reach this academic goal.

Thank Gods, great people who have built everything for us, and my ancestors who may have been blessing me in the haven.

Thank you, everyone.

1.0 Introduction

1.1 Motivations

Complex biological systems are incredibly diverse, ranging from the molecular level to ecosystems. Exploring them allows us to gain fundamental benefits to our health. For instance, elucidating complex biological systems enables us to identify external environmental and lifestyle factors influencing individual health. More importantly, by studying complex biological systems using computational methods and understanding the intricacies of the human body, we can gain insights into the disease and disorder mechanisms and pave the way for the development of new diagnostic tools, treatments, and therapies [1], [2]. This knowledge is also vital for personalized medicine and disease prevention strategies.

Studying complex systems using computational methods can also contribute to the technological development in computational modeling, data analysis, and simulation, which are applicable to a wide range of scientific domains [3]. For instance, understanding brain neurons behaviors gave the inspiration to the creation of artificial neural network models. With these powerful computational methods and predictive models, we can further simulate biological processes and comprehend them even better. Thus, we are entering into a positive-feedback circle, where the more we learn about complex biological systems the more powerful computational methods evolve and vice versa.

Biological research generates vast amounts of data, including genomic, transcriptomic, clinical, and imaging data. Analyzing and interpreting these complex datasets systematically and jointly using computational methods brings down research cost compared to conducting laboratory

experiments [1], [4]. Therefore, studying complex biological systems offers a wide range of benefits, from deepening our understanding of life and evolution to fostering medical advancements and innovative technologies. These motivations contribute to the ever-growing interest and significance of research in this field.

1.2 Problems & Challenges

In general, understanding complex biological systems presents numerous challenges due to their intricate nature and the multitude of interacting factors involved [5]–[10]. Some of the non-technical challenges cannot be addressed using computational methods, like ethical considerations and interdisciplinary collaboration. The ethical consideration challenge arises when we need to involve experiments on living organisms while studying complex biological systems. Addressing these ethical consideration challenges, such as ensuring animal welfare or obtaining informed consent in human studies, poses a complicated challenge. Another non-technical challenge is interdisciplinary collaboration as understanding complex biological systems requires collaboration among scientists from diverse disciplines, including biology, physics, mathematics, computer science, and engineering. While some of the technical challenges include:

a) Integrative Analysis of Diverse Data: Biological research generates vast amounts of data, including genomic, transcriptomic, clinical, and imaging data. Analyzing and interpreting these complex datasets requires sophisticated computational and statistical methods.

b) Data Nonlinearity: Complex biological systems often exhibit nonlinear behavior, meaning that small changes in one entity can lead to disproportionate effects on the

system as a whole. With this nonlinearity, predicting the outcomes of perturbations or interventions in such systems can be challenging.

c) Contextual Causal Factors: Complex biological systems are influenced by multiple interacting factors or contexts, including genetic, environmental, and stochastic elements. Given a certain context, disentangling the contributions of individual factors and understanding their collective impact is a complex task.

d) Subtype Network Comparison: Complex biological systems exhibit significant heterogeneity at various subtype or phenotype networks, for example, different subtypes of a disease mechanism or diverse phenotypes of a cancer. However, it is challenging to identify similarities and differences by comparing these subtype networks. Some technical factors can also pose a challenge while using computational techniques to study biological systems. For example, there are many networks comparison measuring matrices, such as edit distance, degree matrix, and adjacency matrix, where each of them compares two networks in a different aspect (the degree matrix can identify the node-related equivalence of networks and the adjacency matrix can capture the structural equivalence of networks).

e) Causal Mechanism Complexity: Complex biological systems, such as cells and organs, are incredibly complex with numerous interconnected components. The interactions and feedback loops among various elements make it difficult to unravel cause-and-effect relationships and understand the system as a whole, as in the study of causal inference, it is believed a causal mechanism or graph cannot involve a cycle.

Despite these challenges, ongoing advancements in computational modeling and interdisciplinary collaborations are steadily improving our understanding of complex biological systems. For almost all these technical challenges, computational methods can be exploited to solve them. In this thesis, we designed a few computational methods to address some of these challenges, which will be discussed in the following sections.

1.3 Background

Over the years, advancements in ML and DL have revolutionized the way we approach and comprehend complex systems. ML and DL methods have the unique capability to identify patterns, extract meaningful features, and make accurate predictions from complex system data. These techniques leverage the inherent computational power to handle vast amounts of information, enabling researchers to navigate through intricate networks and gain insights that were previously unattainable using traditional analytical approaches.

One of the important applications of ML and DL is the analysis of biomedical data to better understand complex biological systems. With the advent of high-throughput technologies, vast amounts of genetic and protein sequence data are being generated at an unprecedented rate. ML and DL algorithms can learn from the complex biological networks built using these datasets, uncovering hidden relationships between genes, proteins, and biological functions.

Understanding heterogeneities and dynamics among complex biological networks (two of the challenges mentioned above), such as subtype biological networks for a disease, has been a longstanding problem in the field of biology, requiring in-depth knowledge and analysis of intricate interactions between various components. A biological network consists of biological

entities (nodes) and relations (edges) between biological entities within a complex biological system [11]. Biological networks, as any other common real-world networks, appear in many forms or categories, forms like undirected, directed, bidirected, weighted, bipartite, multi-edge, hypergraphs, and trees [12]. Some well-known complex biological networks include protein–protein interaction networks, genetic regulatory networks (DNA–protein interaction networks), metabolic networks, signaling networks, and neuronal networks [12]. As DL is based on artificial neural networks (ANNs) and ANNs are inspired by the biological neural networks [13], thus, DL can be regarded as a form of biological neural networks. By definition, an ANN consists of a collection of artificial neurons inspired by working mechanism of the neurons in a biological brain. Neurons pass wights or signals to the neurons at the next layer, just like the synapses in a biological brain. Then, the receiving neurons process weights or signals and pass the output signals to the neurons at the next level [13].

By harnessing the power of computational algorithms and large-scale data analysis, ML and DL have proven to be invaluable tools in deciphering the inner workings of biological networks at different levels. DL has emerged as a transformative approach to overcome the limitations of traditional ML methods. DL models, such as deep neural networks (DNNs), are capable of automatically learning hierarchical representations directly from raw data, eliminating the need for manual feature engineering. This enables them to capture intricate patterns and dependencies within complex biological systems [14], [15]. DL techniques have shown remarkable success in diverse biological applications, including clinical and library image analysis, genomics [16]–[18].

ML and DL have greatly enhanced our understanding of fundamental biological processes at different levels. Since ML techniques employ statistical models and algorithms to identify

patterns or make predictions based on input data, they have also been successfully applied to various biological problems, such as gene expression analysis, identifying sequence motifs, predicting protein structures, and elucidating gene regulatory networks, and disease diagnosis [15], [19], [20].

In recent years, ML and DL have also played a crucial role in drug discovery, and personalized medicine development [16]–[18]. Traditional methods for identifying potential drug candidates are time-consuming and costly, often resulting in high failure rates. ML and DL models, on the other hand, can rapidly screen large chemical libraries, predict drug-target interactions, and optimize drug properties, leading to more efficient and targeted drug discovery pipelines.

DL models have also been employed to predict protein-protein interactions, classify cancer subtypes, and identify regulatory elements in the genome [14], [16], [17]. These achievements highlight the potential of deep learning to unravel intricate biological phenomena that were previously challenging to decipher.

1.4 Research Statement

Despite their tremendous potential, integrating ML and DL into biological research is not a trivial task. To tackle the challenges and problems mentioned above, specific computational models must be designed. The following are the three projects that we worked on during this program.

1.4.1 Project 1 - 3′-UTR Shortening Contributes to Subtype-Specific Cancer Growth by Breaking Stable ceRNA Crosstalk of Housekeeping Genes

Research Question: In breast cancer, different subtypes of tumor samples, such as estrogen receptor positive and negative (ER+ and ER-), are characterized by distinct molecular mechanisms, including possible differences in the post-transcriptional regulation called the competing-endogenous RNA (RNAs that interact through the competition of microRNA binding). While we can construct the ER+ competing-endogenous RNA (ceRNA) network by applying a traditional correlation cutoff (>=0.6) because there are enough number of samples of ER+ breast cancer (n=77 in the Cancer Genome Atlas database (TCGA)), it is not clear how to identify ER- normal ceRNA network comparable to ER+ ceRNA network because there are only 20 ER- samples in the TCGA (n=20). **Challenges**: We found that ER+ and ER- subtypes provide different sample sizes of samples (n=77 and 20, respectively), biasing the ceRNA network size and disabling the fair comparison of the network dynamics. And it is not straightforward to identify ER- normal ceRNA network comparable to ER+ normal ceRNA network.

Novelties, Implementation Difficulties, Existing Methods, and Advantages of Our Method: To shed a systematic understanding between two biological conditions, one can compare network models that represent the conditions. However, there is no dedicated algorithm to identify comparable networks from the conditions of different sample sizes. Due to the absence of such methods, we tried several straightforward methods to address these challenges and identify the ER- ceRNA network comparable to ER+ ceRNA network. Below are the results we obtained from the experiments. First, the same cutoff will inflate the number of edges for the ER- network. Second, subsampling the ER+ normal samples to match the number of samples for ER- (n=20) does not work because the subsampled ceRNA networks do not keep topological consistency

among themselves. Third, we could not use the co-expression cutoff that makes the same statistical significance to ER+, because, to achieve the same statistical significance of the traditional cutoff value (0.6) of ER+, the cutoff value of ER- would inflate to 0.91 which results in a drastically deflated number of edges. To address this and identify comparable networks between ER+ and ER- breast cancer, our method leverages a network similarity measure and studies its performance with extensive experiments. Consequently, this method could identify the ceRNA network from ER- samples that demonstrated a similar structure and properties of the graph to ER+ network.

1.4.2 Project 2 - Deep Neural Networks With Knockoff Features Identify Nonlinear Causal Relations and Estimate Effect Sizes in Complex Biological Systems

Research Questions: Complex biological systems are characterized by non-linear associations [21], [22]. For example, the effects of hormone receptor status on breast cancer biology are often nonlinear due to their complex interactions with other molecular complexes in multiple regulation processes[23]–[25]. Another example of nonlinearity in biological systems is how molecular/clinical features are interacting for patients' phenotypes (e.g., clinical outcomes). When they interact, they often are regulated through multiple biochemical pathways[26], and thus these relations are likely nonlinear. Thus, the question is how to model and capture the nonlinearity in computational analyses of biological systems. **Challenges**: Because of this nonlinearity, learning causalities and estimating the effect size in such systems can be challenging. Learning causal relationships between clinical features is vital for making informed medical decisions, developing effective treatments, and improving overall patient outcomes. Thus, to allow for learning nonlinear causal relationships between clinical/molecular features, it is crucial to develop a method that can learn nonlinear causal relationships. Additionally, estimating effect size in causal

inference is critical for understanding the practical significance and implications of research findings. For example, it aids in making informed decisions, comparing interventions, and planning future studies to advance our understanding of causal relationships. Thus, it is also crucial to enable the estimation of the effect size especially in the nonlinear causal relationships.

Novelties, Implementation Difficulties, Existing Methods, and Advantages of Our Method: Our method, causal Directed Acyclic Graphs using deep-learning VAriable SElection (DAG-deepVASE), explicitly learns nonlinear causal relationships, linear causal relationships, and estimate their effect sizes using a deep-neural network approach coupled with the knockoff framework. This method involves many computational techniques such as Mixed Graphic Model and knockoff data generation. Previously, causal inference has been approached using traditional fashion, either constraint-based or score-based, and deep-learning based searching. Peter and Clark (PC) [27], one of the most popular algorithms under the constraint-based approaches [28]–[33], runs in two steps. The first step is to use a combination of conditional independence tests; in the second step, it uses graph pruning techniques to determine a skeleton of the directed acyclic graph and then to determine the causal directions in the skeleton network. PC usually produces many bidirectional causal relations because it is constrained-based causal inference method where it uses rules to determine causal directions. Under the score-based approach, Degenerate Gaussian (DG) [34] is a recently proposed method extending the widely used likelihood score function BIC score [35], [36]. It was designed for processing mixed types of data by embedding discrete variables into a continuous space using one hot vector representation. On the other hand, several deep-learning approaches have been proposed. DAG-GNN [37], proposed by Yu et al., is a deep generative model and applies a variant of the structural constraint to learn the directed acyclic. Zheng et al. extended DAG-GNN by generalizing it so various approximations can be used for search

(NOTEARS) [38]. Although deep-learning approaches successfully generalized the problem of causal inference and facilitates the use of advanced deep-learning techniques for this problem, these two methods are limited due to the common strategy that they search over large directed cyclic graphs (DAGs). They only use the DNN component to address nonlinearity in how DAGs are searched through, not to address nonlinearity in each relationship. On the contrary, DAG-deepVASE utilizes the deep learning component in a completely novel way to explicitly learn both nonlinear causal relationships, as well as their effect sizes. In our method, we explicitly set one of the features as the target and identify related features in each iteration while learning nonlinear associations. This explicit setting enables us to identify nonlinearity in each relationship.

1.4.3 Project 3 - Deep Neural Network Jointly Learning Gene Expression and Biological Condition Information Identifies Cell Subtypes Nonlinearly Linked to the Biological Condition

Research Questions: Among many cell types, biologists and clinicians are usually interested in cell types that are related to a particular biological condition, e.g., a particular pathological state among multiple phases of a disease (disease-specific cell subtypes). The question is how to exploit this biological condition information to better identify condition-specific cell subtypes. **Challenges**: Existing methods only use gene expression data to identify cell types and thus do not fully leverage the biological condition information to identify condition-specific cell subtypes. The challenge is to develop a method that leverages not only gene expression information but also biological condition information to accurately identify condition-specific cell types.

Novelties, Implementation Difficulties, Existing Methods, and Advantages of Our Method: While existing methods, like Leiden, Louvain, and Milo, only use gene expression information to identify clusters, our method jointly learns cell types using gene expression, a biological condition, a set of known cell types to accurately identify finer cellular states linked to a biological condition with the highest sensitivity and specificity. Since our method is based on DNN, compared with other clustering methods, our method can capture nonlinear cell dynamics that other methods cannot. Louvain and Leiden are two of the most widely used clustering methods in computational biology. Louvain aims to detect communities in complex biological networks based on a modularity score. The modularity score quantifies the quality of an assignment of nodes to communities, therefore, it tries to maximize a modularity score for each community. Leiden is an extension of Louvain where it can find some communities where Louvain finds them not well-connected. Milo is a method for differential abundance analysis on a K-nearest Neighbor (KNN) graph from single-cell RNA sequencing data. Since these methods only use gene expression, they are not designed to identify the cell types that are abundant in related biological conditions due to three limitations: 1) it does not consider the impact of one criterion on another; samples of a particular biological condition would render distinct biological functions represented with distinct molecular behavior and this distinct molecular behavior can have an impact to the level of gene expression; 2) it disregards the dimensional differences in the criteria, as the gene expression information typically represents several thousand genes while only one particular type of biological condition is exploited at any given time; and 3) the optimizations rely on linear modeling that they do not capture the nonlinear relationships between cell identity and gene expression/biological condition information. On the other hand, our method addresses all the limitations. First, our method simultaneously optimizes two loss functions, L_t for the gene-

expression information and L_z for the biological condition information to identify the optimal solution in terms of both the biological condition and gene expression. Second, in identifying the solution, we balance the weights between the high dimensional gene expression data and the biological condition information by controlling the weight of the gene expression information vs. that of the biological condition information. Third, to capture the nonlinear relationships, we utilize a DNN component to encapsulate the complex and nonlinear relationships using multiple layers of nonlinear activators.

1.5 Contributions

1.5.1 Contributions of Project 1 "3′-UTR Shortening Contributes to Subtype-Specific Cancer Growth by Breaking Stable ceRNA Crosstalk of Housekeeping Genes"

Contributions to Computational Biology: With our method, we made a discovery that would be impossible without it that house keeping (HK) genes can play a novel role as stable and strong miRNA sponges (sponge HK genes) that synchronize the ceRNA networks of normal samples (adjacent to ER+ and ER– tumor samples). We also identified 3′US events (3′untranslated region (UTR) shortening that removes microRNA binding sites located in the 3′UTR of genes) in the ER- tumor break the stable sponge effect of HK genes in a subtype-specific fashion, especially in association with the aggressive and metastatic phenotypes. Our findings bring a new perspective on the role of previously unexplored class of genes, house keeping genes on the breast cancer etiology. **Contributions to Computer Science**: To address the challenge and identify the ER- network comparable to ER+ network, we proposed a systematic way utilizing a network similarity

measure called normalized Laplacian matrix eigenvalue distribution. We exploited various network analysis techniques to study heterogeneities and dynamics differentiating the breast cancer subtypes. Furthermore, through our findings, we demonstrated that this network comparison method successfully brings biologically meaningful and innovative findings.

1.5.2 Contributions of Project 2 "Deep Neural Networks With Knockoff Features Identify Nonlinear Causal Relations and Estimate Effect Sizes in Complex Biological Systems"

We developed a causal inference method that learns both linear and nonlinear causal relations and estimates the effect size using a deep-neural network approach coupled with the knockoff framework. The DNN approach allows for identifying nonlinear relationships between the input features and the knockoff framework allows for estimating the effect size of the associations between the input features. Our method outperforms existing methods in identifying true and known nonlinear causal relations. The identified nonlinear causal relations and estimating their effect size can help understand the complex disease pathobiology, such as breast cancer, pediatric sepsis, and the effect of gut microbiome on BMI, which is not possible using other methods. We created a GitHub repository (https://github.com/ZhenjiangFan/DAG-deepVASE) to make this causal inference method public so that computational biologists can utilize and extend our method.

1.5.3 Contributions of Project 3 "Deep Neural Network Jointly Learning Gene Expression and Biological Condition Information Identifies Cell Subtypes Nonlinearly Linked to the Biological Condition"

Contributions to Computational Biology: As integrated biomedical datasets like multi-omics data are receiving more attention from researchers, our computational model creates a new direction for understanding complex biological systems by analyzing multiple data resources because researchers can follow our DNN model design where the model takes multiple data resources (data resources collected from the same set of samples) as input, such as mRNA expression, DNA methylation, and microRNA (miRNA) expression. Our approach could also inspire other computational biologists to start thinking about utilizing other biological information (e.g., cell spatial information) while building their DNN model. **Contributions to Computer Science**: We designed and implemented a DNN-based joint-learning method that simultaneously optimizes multiple data sources of different nature (e.g., gene expression, a biological condition, a set of known cell types). In incorporating the data sources, our method explicitly leverages the different levels of method maturity for each source. For example, while clustering methods using gene expression information, like Leiden, Louvain, and Milo are well developed, there is no method either using the biological condition information or incorporating the information to gene expression information to accurately identify condition-specific cell types (e.g., cellular states linked to a biological condition). By first embedding the clustering results from such a method that is based on gene expression information and jointly training the biological condition information on the embedding, we achieve the better sensitivity and specificity than other methods that use only gene expression information. Generally, many real-world datasets need to be interpreted jointly with other data. Therefore, our method can also be utilized as a general clustering method

in such a scenario. We also created a GitHub repository for this method and made it public for researchers to use, which can be found at https://github.com/ZhenjiangFan/scDeepJointClust.

1.6 Future Work

1.6.1 Future Work for Project 1 "3′-UTR Shortening Contributes to Subtype-Specific Cancer Growth by Breaking Stable ceRNA Crosstalk of Housekeeping Genes"

Our first project aimed to study the distinct ceRNA dynamics between the ER+ and ER- group of tumor samples. Although ER status is an important clinical variable [39], it is important to note that the two groups do not directly represent further clinical subtypes of breast cancers, such as HER2+ or Triple-Negative. Thus, to reveal further clinical relevance of the ceRNA dynamics, more study is warranted in direct clinical subtypes within each group.

1.6.2 Future Work for Project 2 "Deep Neural Networks With Knockoff Features Identify Nonlinear Causal Relations and Estimate Effect Sizes in Complex Biological Systems"

DAG-deepVASE clearly has many advantages over existing methods. However, our method can be improved in two aspects. The first aspect is that our method cannot take nonordinal categorical variables as the knockoff generation approach use in this work, model-X knockoff, assumes Gaussian distribution, whereas nonordinal variables do not follow Gaussian distribution. Therefore, one future work would be generating the knockoff variables for nonordinal categorical

variables based on a regression model for nonordinal categorical variables [40]. The second future work could focus on finding a way to estimate statistical significance of the likelihood ratio test we derived to determine the causal direction as our method currently only estimate the effect size but not statistical significance.

1.6.3 Future Work for Project 3 "Deep Neural Network Jointly Learning Gene Expression and Biological Condition Information Identifies Cell Subtypes Nonlinearly Linked to the Biological Condition"

Our approach capitalizes on the fact that condition-specific cells show an enrichment of a specific biological condition in the cells of the same type. Thus, our approach may not be useful to identify cell types that do not exhibit such an enrichment pattern. Therefore, a future work may increase the sensitivity on the enrichment degree by testing weights to the DNN model to adjust the importance of the biological conditions compared to the gene expression information. Another direction for future work concerns identifying cell subtypes or cell states of the same type. Currently, our method is to refine the cell type definition constructed based on the gene expression with the biological condition information. However, the cell types may be further divided based on the enrichment of the biological conditions. For example, a cell type can have multiple states that are differentially enriched in biological conditions. Then, it is interesting to identify such cell states since they can inform further treatment strategies. In that sense, another future work may aim to divide the number of cell types identified using the gene expression information. (e.g., cluster number (K)) so the divided cell types can represent different cell states of the same type that show different enrichment pattern to the biological condition. For this, we would need to build another model that has three components, an evaluation estimator (e.g., K-fold cross-validation),

a clustering performance evaluation matrix (e.g., Silhouette score), and a custom loss function to the DNN model.

2.0 Project 1 - 3′-UTR Shortening Contributes to Subtype-Specific Cancer Growth by Breaking Stable ceRNA Crosstalk of Housekeeping Genes

2.1 Summary

Shortening of 3′UTRs (3′US) through alternative polyadenylation (APA) is a post-transcriptional mechanism that regulate expression of hundreds of genes in human cancers. In breast cancer, different subtypes of tumor samples, such as estrogen receptor positive and negative (ER+ and ER-), are characterized by distinct molecular mechanisms, suggesting possible differences in the post-transcriptional regulation between the subtype tumors. In this study, based on the profound tumorigenic role of 3′US interacting with competing-endogenous RNA (ceRNA) network (3′US-ceRNA effect), we hypothesize that the 3′US-ceRNA effect drives subtype-specific tumor growth. However, we found that the subtypes are available in different sample size, biasing the ceRNA network size and disabling the fair comparison of the 3′US-ceRNA effect. Using normalized Laplacian Matrix Eigenvalue Distribution, we addressed this bias and built the tumor ceRNA networks comparable between the subtypes. Based on the comparison, we identified a novel role of housekeeping (HK) genes as stable and strong miRNA sponges (sponge HK genes) that synchronize the ceRNA networks of normal samples (adjacent to ER+ and ER- tumor samples). We further found that distinct 3′US events in the ER- tumor break the stable sponge effect of HK genes in a subtype-specific fashion, especially in association with the aggressive and

metastatic phenotypes. Knockdown of NUDT21 further suggested the role of 3′US-ceRNA effect repressing HK genes for tumor growth. In this study, we identified 3′US-ceRNA effect on the sponge HK genes for subtype-specific growth of ER- tumors.

2.2 Introduction

Approximately, 70% of human genes contain multiple polyadenylation (polyA) sites in the 3′-untranslated region (3′-UTR) [41]. Through alternative polyadenylation (APA) during transcription, messenger RNAs (mRNA) from the same gene can have various 3′-UTR lengths. Since the 3′-UTR contains regulatory regions including microRNA (miRNA) target sites, mRNAs with shortened or lengthened 3′-UTRs may diversify the regulation landscape, for example miRNA binding landscape. In human cancer, 3′-UTR lengthening (3′UL) has been associated with cell senescence [42] with implications for tumor-associated processes, such as cell cycle inhibition, DNA damage/repair process, and tumor suppression [43]–[46]. Widespread 3′-UTR shortening (3′US) has been reported for diverse types of human cancer [41]. Further, 3′US events add prognostic power beyond common clinical and molecular covariates in cancer patients [47] and are associated with drug sensitivity in cancer cell lines [48]. These results suggest that APA events, both 3′-UTR shortening and lengthening, play important roles in cancer etiology and treatments.

The 3′-UTR is also implicated in competing-endogenous RNA crosstalk (ceRNA) [49]. CeRNAs co-regulate each other RNAs through competing for binding miRNAs. In diverse types of cancer, ceRNA regulation involves established oncogenes and tumor suppressor genes [50] and facilitates molecular pathway interactions for tumorigenesis [51]. When 3′-UTR shortening genes lose miRNA target sites on their 3′-UTRs and do not sequester the miRNAs, the associated

miRNAs bind to the 3′-UTR of the ceRNA partners. As a result, 3′-UTR shortening disrupts ceRNA crosstalk (3′US-ceRNA effect) for growth in diverse types of cancer, including breast cancer [52]. In a recent study, we showed that this 3′US-ceRNA effect promotes tumor growth independent of potential confounding factors, such as somatic mutation status (SNPs and small INDELs), tumor purity, immune cell infiltration, cell proliferation, or miRNA biogenesis and expression [53].

Breast cancer can be classified into two major subtypes based on the presence or absence of estrogen receptor (ER) [39]. Estrogen receptor positive (ER+) breast tumors grow in the presence of the hormone estrogen. So, ER+ cancers can be treated with endocrine therapy which blocks ER activity or depletes estrogen levels. On the other hand, estrogen receptor negative (ER-) breast tumors have unique growth mechanism due to absence of the estrogen receptor. The unique growth mechanism of ER- tumors makes it difficult to treat ER- breast cancer that has a worse prognosis than ER+ [54] with a more aggressive phenotype [55], [56]. Based on the profound tumorigenic effect of 3′US-ceRNA [52], we hypothesize that 3′US-ceRNA effects specific to ER- breast tumors contribute to the unique growth mechanism. In this study, we tested this hypothesis by addressing a quantitative challenge due to different sample sizes between ER+ and ER- breast tumor samples. As a result, we identified a novel subset of housekeeping (HK) genes (sponge HK) effectively sponging miRNAs to synchronize the ceRNA networks in normal samples (adjacent to the subtype tumor samples). Further, we showed that the 3′US-ceRNA effects repress the sponge HK genes, leading to subtype-specific tumor growth. In ER- breast tumor, this subtype-specific tumor growth is associated with aggressive and metastatic phenotypes of ER- tumors, attributing its unique grow mechanism partially to subtype-specific 3′US-ceRNA effects.

2.3 Materials and Methods

2.3.1 TCGA Breast Tumor RNA-seq Data and Identification of Breast Cancer Subtypes

Quantified gene expression files (RNASeqV1) for primary breast tumors (TCGA sample code 01) and their matching solid normal samples (TCGA sample code 11) were downloaded from the TCGA Data Portal[57]. We used 97 breast tumor samples that have matched normal tissues, which were further categorized into 77 estrogen receptor positive (ER+) and 20 estrogen receptor negative (ER-). For ER+ and ER-, we collected both normal (ER+ normal and ER- normal) and tumor (ER+ tumor and ER- tumor) samples. A total of 10,868 expressed RefSeq genes (fragments per kilobase of transcript per million mapped reads (FPM) ≥ 1 in $> 80\%$ of all samples) were selected for downstream analyses.

2.3.2 Selection of miRNA Target Sites

Predicted miRNA-target sites were obtained from TargetScanHuman version 6.2 [58]. Only those with a preferentially conserved targeting score (Pct) more than 0 were used [47]. Experimentally validated miRNA- target sites were obtained from TarBase version 5.0 [59], miRecords version 4 [60] and miRTarBase version 4.5 [61]. The target sites found in indirect studies such as microarray experiments and high-throughput proteomics measurements were filtered out [62]. Another source is the microRNA target atlas composed of public AGO-CLIP data [63] with significant target sites (q-value < 0.05). The predicted and validated target site information was then combined to use in this study. Among 1,261 miRNAs curated in the TCGA

BRCA data, we used 713 expressed ones (avg. FPM > 1) in our analyses (**Supplemental Table. 1. Tab 7**).

2.3.3 Statistical Significance of Pearson Correlation Coefficient

The implementation of the Pearson r function is provided by a python package, SciPy, and available at https://scipy.org/, which returns the calculated correlation coefficient and a 2-tailed p-value for testing non-correlation. The Pearson correlation coefficient measures the linear relationship between two variables (e.g. gene X and gene Y) and when the two covariates follow binormal distribution, we can assume that their Pearson's correlation follows student t distribution. The p-value is calculated by three steps: 1) calculating the value of the Pearson's correlation *t*, 2) defining the degree of freedom *df* (*N-2*, where *N* is the sample size), 3) getting the probability of having *t* or more extreme than *t* from a Student's t-distribution with the degrees of freedom *df*. We used hypergeometric test in Scipy to estimate significant of miRNA binding site overlap between genes.

2.3.4 Detection of APA Events

We used DaPars [47] to identify 3′UTR shortening and lengthening in RNA-Seq data based on the same cutoff and parameter values optimized in the original paper. We checked that our prediction is 100% matched with that of the original DaPars result. The DaPars paper provided multiple lines of evidence to demonstrate that DaPars indeed identified APA events in the TCGA data. First, 51% of the DaPars predictions are within 50 bp of the annotated APAs compiled from Refseq, ENSEMBL, UCSC gene models and polyA database (polyA_DB[64]). Second, in the

upstream (-50 nt) of the predicted APA sites, MEME motif enrichment analysis [65] successfully identified canonical polyA signal AATAAA.

2.3.5 Housekeeping, Transcription Factor and Tumor-Associated Genes

Housekeeping genes are required for the maintenance of basic cellular functions that are essential for the existence of a cell, regardless of its specific role in the tissue or organism. Generally, housekeeping (HK) genes are expected to be expressed at relatively constant rates in most non-pathological situations [66]. We used 3,804 HK genes defined in RNA-Seq data for 16 normal human tissue types: adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells [67].

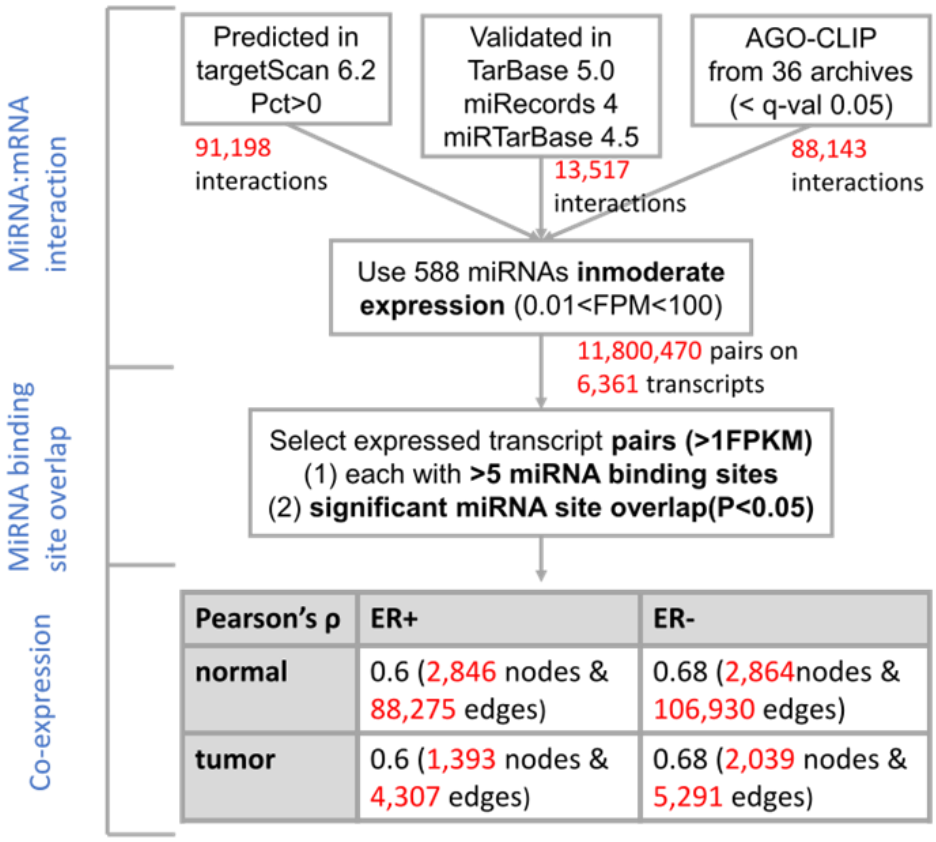
Transcription factors (TFs) play an important role in the gene regulatory network. We downloaded 2,020 TF genes defined in TFcheckpoint database [68], in which TF information is collected from 9 different resources. Among them, we used 1,020 genes that are further supported by sequence-specific DNA-binding RNA polymerase II activity.

The tumor-suppressor genes and oncogenes were defined by the TUSON algorithm from genome sequencing of > 8,200 tumor/normal pairs[69], in particular residue-specific activating mutations for oncogenes and discrete inactivating mutations for tumor-suppressor genes. TUSON computationally analyzes patterns of mutation in tumors and predicts the likelihood that any individual gene functions as a tumor- suppressor gene or oncogene. We used 466 oncogenes and 466 tumor suppressor genes at the top 500 in each prediction (after subtracting 34 genes in common).

2.3.6 Building Subtype ceRNA Networks

For each of the breast cancer data (ER+ normal, ER+ tumor, ER- normal, and ER- tumor) that we defined above, we constructed a ceRNA network based on microRNA (miRNA) target site share and expression correlation[52], [70]. The same miRNA target site information was determined regardless of the subtypes, resulting into the miRNA target site share network (based on FDR > 0.05 in hypergeometric test with miRNA target site information). And given the same miRNA target site share network, the expression correlation information for each subtype will select ceRNA network edges for each subtype.

We first constructed the ER+ normal reference ceRNA network by applying a traditional correlation cutoff (>=0.6) on the miRNA target site share network. Then, to identify ER- normal ceRNA network comparable to ER+ normal reference ceRNA network, we applied different correlation cutoff values (0 to 1 with a step size of 0.01) on the miRNA target site share network for ER- normal samples, and select the correlation cutoff values that makes ER- normal ceRNA network most similar to ER+ normal reference ceRNA network. To estimate topological similarity, we employed normalized Laplacian Matrix Eigenvalue Distribution that discovers ensembles of Erdős–Rényi graphs better than other metrics such as Sequential Adjacency or Laplacian[71]. After identifying the ER+ normal reference network and the corresponding ER- normal network, we used the same cutoffs (0.6 for ER+ subtypes and 0.68 for ER- subtypes) to construct the ER+ tumor network and the ER- tumor network, respectively. An overall workflow is in **Supplemental Figure. 1**.



Supplemental Figure 1. Workflow for the ceRNA network construction for the TCGA breast tumor and the matched normal samples of ER+ and ER- subtypes.

2.3.7 Estimating Topological Similarity

To identify the structural equivalence between two networks, we employed spectral analysis not only to identify the structural similarities, but also to track down the underlying dynamic behavior changes between them. Spectral clustering on networks uses the eigenvalues of several matrices, such as adjacency matrix, the Laplacian matrix, the normalized Laplacian matrix. In this research, we used the normalized Laplacian matrix since it involves both the degree matrix and adjacency matrix, where the degree matrix can identify the node related equivalence of networks and the adjacency matrix can capture the structural equivalence of networks. Another

very important reason of using the normalized Laplacian eigenvalue matrix is that it is more sensitive to small changes because it considers more information.

For network G, the normalized Laplacian of G is the matrix:

$$N = D^{-1/2} - LD^{-1/2} \tag{1}$$

where L is the Laplacian matrix of G and D is the degree matrix. The Laplacian matrix L is defined as: $L = D - A$, where A is the adjacency matrix of G.

In N, each of its entry elements is given by:

$$N_{i,j} = \begin{cases} 1, & if\ i = j\ and\ \text{degree}(v_i) \neq 0 \\ -\dfrac{1}{\sqrt{\text{degree}(v_i)\ \text{degree}\ (v_j)}}, & if\ i \neq j\ and\ v_i\ is\ adjacent\ to\ v_j \\ 0, & otherwise \end{cases} \tag{2}$$

where degree (vertex v) is the function that returns the degree of the vertex v.

To assess how close two network G_1 and G_2 are, we first built N1 and N2 based on the connection information of G1 and G2, respectively. Then, we defined $dist_1$ and $dist_2$ as the eigenvalue distribution of N1 and N2, respectively. We further used the Kolmogorov–Smirnov test (KS test), which is defined as:

$$K_{1,2} = \sup_x |dist_1(x) - dist_2(x)| \tag{4}$$

where $\sup_x$ is the supremum of the set of distances.

By using the normalized Laplacian Matrix and KS test, ER+ normal reference network G_{ref}^{ER+} is compared with a ER- normal subnetwork with a particular correlation cutoff i G_i^{ER-} in the following three steps:

1) Compute the normalized Laplacian metrics N_{ref}^{ER+} and N_i^{ER-} from G_{ref}^{ER+} and G_i^{ER-} respectively.
2) Compute the eigenvalues E_{ref}^{ER+} and E_i^{ER-} from N_{ref}^{ER+} and N_i^{ER-} respectively.
3) Compute the KS statistic between E_{ref}^{ER+} and E_i^{ER-}.

The third step test the null hypothesis that eigenvalues E_{ref}^{ER+} and E_i^{ER-} are drawn from the same continuous distribution. If the two-tailed p-value returned by the KS test is high, then we cannot reject the hypothesis that G_{ref}^{ER+} and G_i^{ER-} are the same network. In another word, the higher the p-value is, the more similar G_{ref}^{ER+} and G_i^{ER-}.

2.4 Results

2.4.1 Widespread 3′-UTR Shortening and Lengthening Events for ER+ and ER-

To identify subtype-specific APA genes, we first identified 77 ER-positive (ER+) and 20 ER-negative (ER-) sample pairs (breast tumor and the adjacent normal samples) from 97 sample pairs available in TCGA (see Methods). Then, we identified 3′UTR shortened (3′US) and 3′UTR lengthened (3′UL) genes (tumor vs. normal) using DaPars [47] in each subtype. We found that the ER+ and ER- sample pairs have similar numbers of total 3′US genes and 3′UL genes (**Figure. 2.1A**). However, 3′US genes are more recurrent (occurring in > 20% of the tumor samples [47]) in both the subtype tumors (**Figure. 2.1B, C** e.g. $P=5.0\times10^{-5}$ for ER+). Further analyses showed

that 3′US and 3′UL play distinct roles in the subtypes. First, the recurrent 3′US and 3′UL genes show little overlap (1 and 13 genes in common, $P=1.27e^{-6}$ and $P=3.97e^{-9}$, respectively, **Figure. 2.1B, C**). Second, the number of 3′UL events is not correlated with that of 3′US events across the tumor samples (P=0.35 for ER+ and P=0.61 for ER-, **Figure. 2.1D, E**). Third, Ingenuity Pathway Analysis (IPA) shows that the recurrent 3′US and 3′UL genes are enriched for distinct sets of molecular pathways (**Supplemental Table. 1. Tab 1**, **Supplemental Figure. 2**). The IPA analysis further suggests that 3′UL or 3′US genes themselves have limited roles for cancer overall, since a small number of pathways are significantly ($P<10^{-2}$) enriched for them (12 and 14 for 3′UL in ER-/+ and 29 and 3 for 3′US in ER-/+ samples)and at most a couple of them are "cancer" pathways (one for 3′UL in ER+ and two for 3′US in ER- with keyword "cancer"). Based on the profound tumorigenic role of 3′US in its interaction with ceRNAs (3′US-ceRNA effect) [52], we hypothesize that 3′US-ceRNA effect, not 3′US cis effect, promotes ER- specific tumor growth.

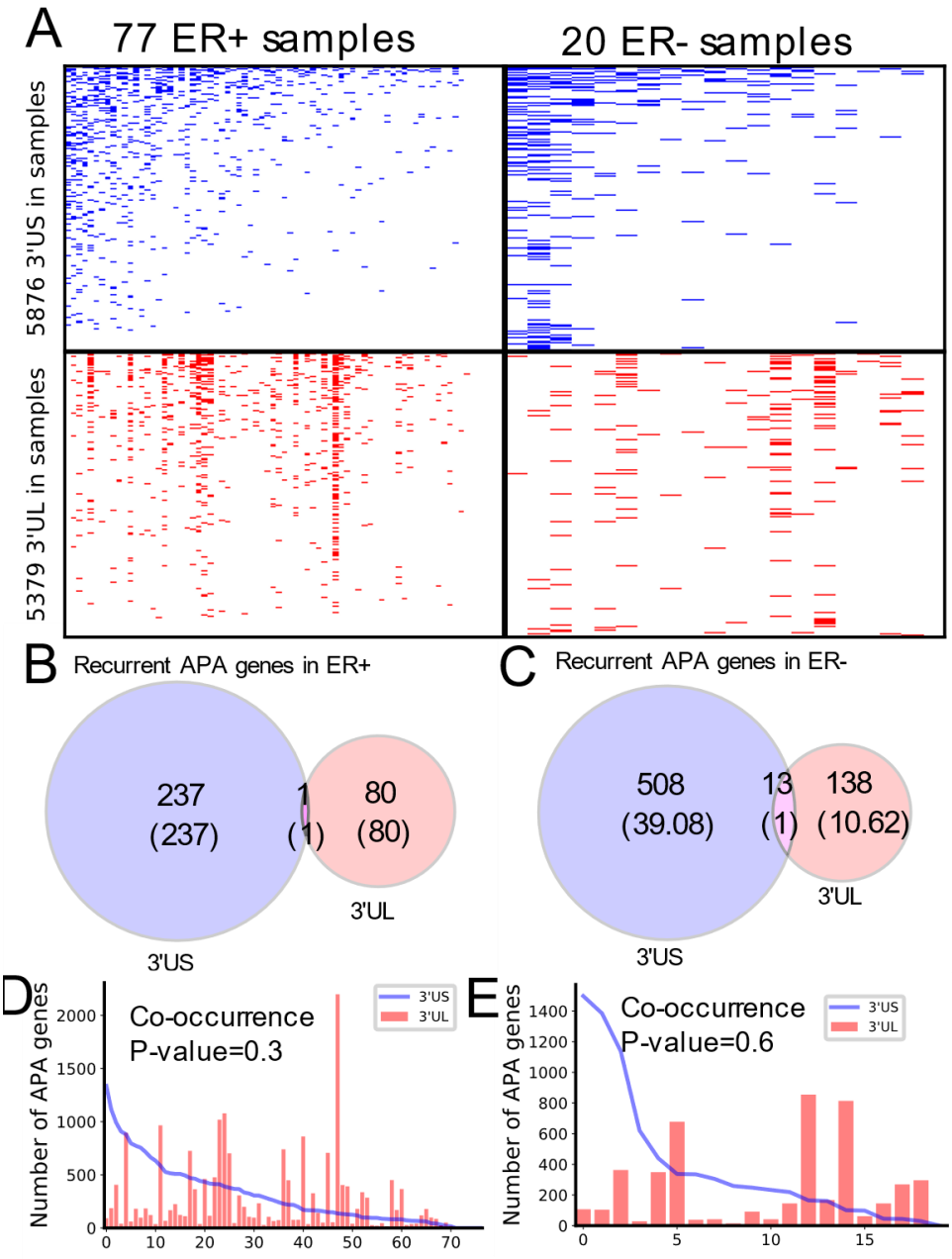
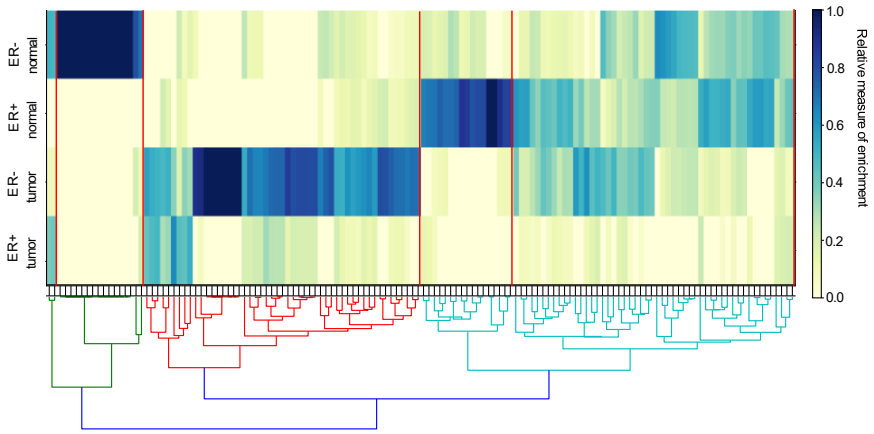


Figure 2.1. A Global APA events distinct for ER+ and ER-. (A). Heatmaps showing the genes with 3′US (top panel) or 3′UL (bottom panel) in ER+ samples (left column) or ER- samples (right column), ranked by the total number of APA events. (B), (C) Overlap of the recurring (>20% in samples) 3′US and 3′UL genes in ER+ and ER-, respectively. (D), (E), The number of APA genes (3′US in line and 3′UL in red bar) in the tumor-normal sample pairs in In ER+ and ER-, respectively, ordered as in Figure. 1.1A.



Supplemental Figure 2. IPA pathways enriched for the recurrent 3′UL and 3′US genes in ER- and ER+. Colors represent enrichment of each pathway (column) for each class of genes (row). The red lines in the heatmap cut the pathways into 5 clusters in accordance with the dendrogram drawn on the bottom.

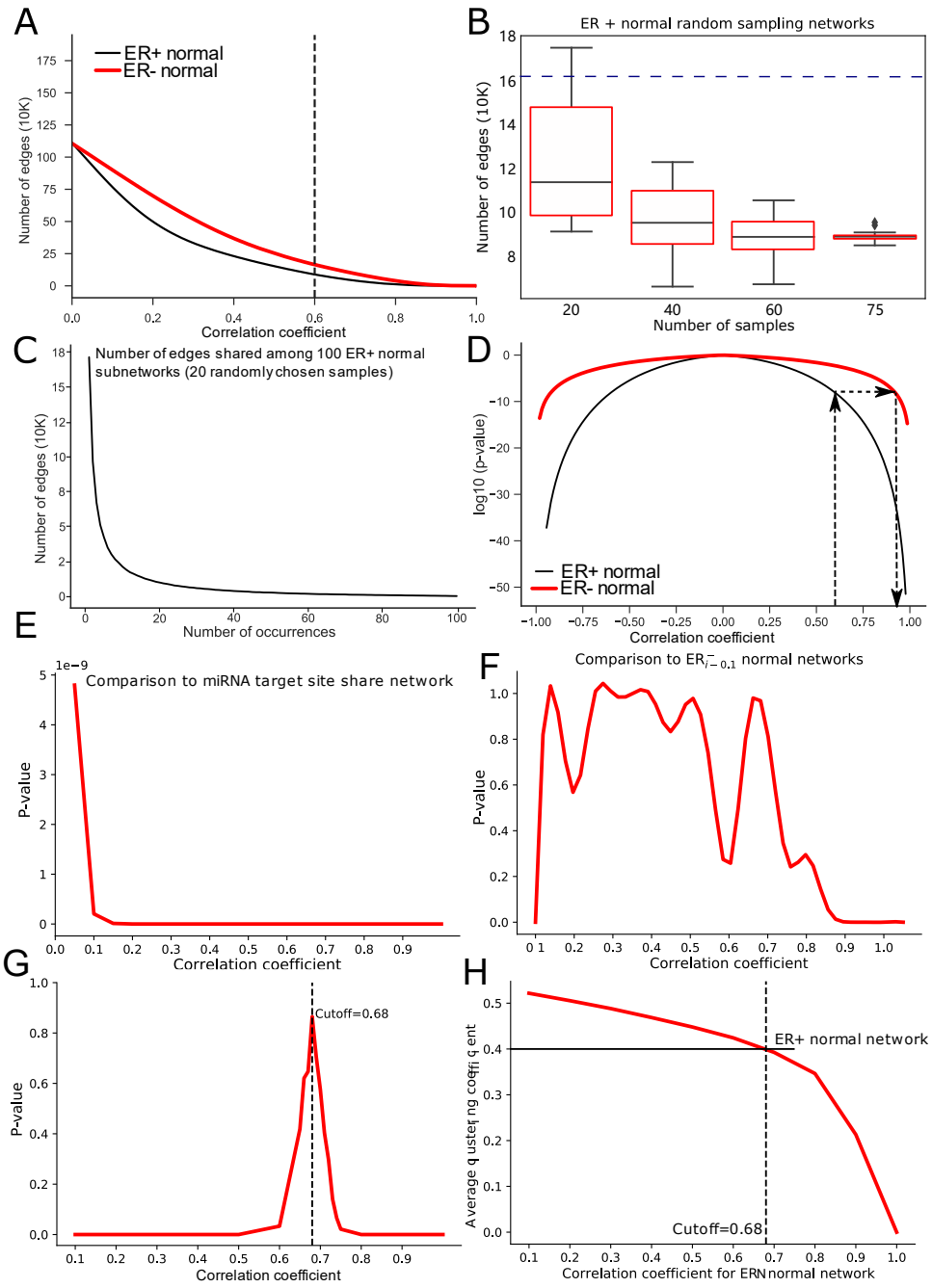
2.4.2 Two-Step Pairwise Normalization of ER+ and ER- ceRNA Network

We previously identified the 3′US-ceRNA effect in the ceRNA network [52]. To identify the 3′US-ceRNA effect specific to ER- tumors, we aim to build ceRNA networks for ER- and ER+ tumors and compare them. Computationally, ceRNA gene pairs in the networks are those that share a significant number of miRNA target sites and are co-expressed [52], [70]. However, using the common co-expression cutoff (e.g., Pearson's $\rho > 0.6$) will inflate the number of edges for ER- (160,687 in ER- normal vs. 88,275 in ER+ normal, **Supplemental Figure. 3A**). To test if this inflation is attributable to the sample size difference, we built the ceRNA network 100 times from different numbers of (20, 40, 60, and 75) normal subsamples from ER+ tumors based on the same co-expression cutoff (**Supplemental Figure. 3B)**. In general, the number of edges in the ceRNA networks increases as the subsample size decreases. Especially, when the same number of samples

(20) to that of ER- normal network is used, the number of edges in the subsampled networks becomes closer to the case of ER- normal network.

Since the network size difference is attributable to the sample size difference, one might want to subsample ER+ normal samples to match the number of samples for ER- (n=20). To assess this solution, we subsampled 20 ER+ normal samples 100 times, built a ceRNA network for each subsample, and collected all the edges (916,999) across the networks. Then, we checked how many times each edge occurs across the 100 subsampled networks. We found that the subsampled ceRNA networks do not keep topological consistency within them, as 22.1% (202,997) of the edges are shared by less than the 20 ceRNA networks (**Supplemental Figure. 3C**). Then, one might want to build the ER- ceRNA network using the co-expression cutoff with the same statistical significance to ER+ (0.91, $P\sim=10^{-8.2}$, **Supplemental Figure. 3D**). To achieve the same statistical significance of the traditional cutoff value (0.6) of ER+, the cutoff value of ER- would inflate to 0.91, resulting in a drastically deflated number of edges (**Supplemental Figure. 3D)**. We addressed this issue in the following way. First, we built the reference network from normal samples of larger size (ER+) using the common correlation cutoff (Pearson's $\rho > 0.6$). Since normal samples should have similar molecular dynamics between ER+ and ER-, we sought to find the co-expression cutoff for ER- normal network that yields the most topological similarity to the ER+ reference network. To estimate topological similarity, we employed a normalized Laplacian Matrix Eigenvalue Distribution that discovers ensembles of Erdős–Rényi graphs better than other metrics, such as Sequential Adjacency or Laplacian [71] (see Methods). While ER- normal network topology changes drastically if different correlation cutoff values are used (**Supplemental Figure. 3E, 3F**), we found that the cutoff 0.68 makes the ER- normal network most similar to the ER+ reference network (**Supplemental Figure. 3G**). Using another measure for topological

similarity, average clustering coefficient [72], the cutoff of 0.68 is supported again since normal ER- network with correlation cutoff 0.68 makes the closest average clustering coefficient to the reference network (0.4, **Supplemental Figure. 3H**). Since normal and tumor ceRNA networks within each subtype share the same number of samples thus would not suffer from this bias [52], [73]–[75], we applied the subtype-specific cutoffs (0.68 for ER- and 0.6 for ER+) to build the tumor ceRNA networks in each subtype.



Supplemental Figure 3. Two-step Pairwise Normalization of ER+ and ER- ceRNA network. (A) Number of edges in the ceRNA networks by the correlation coefficient cutoff (black and red line for ER+ and ER- normal networks, respectively). (B) Number of edges in 100 networks built from a subset of ER+ normal samples in different size. Blue dotted line indicates the number of edges of ER- normal network whose sample size is 20 (160,687) (C) The number of edges shared among 100 ER+ normal samples, where each of them was built by using 20 randomly chosen samples. (D) Statistical significance (p-value) achievable by using different

correlation coefficient cutoff values for ER+ (black) and ER- (red) samples. Statistical significance for a correlation coefficient cutoff value is described in Methods.. (E) Comparison of ER- normal network with the miRNA target site share network to by correlation cutoff value (see Methods). (F) Comparison of ER- normal network with that of the previous correlation cutoff value in the stepwise increase (see Methods). (G). Significance of topological similarity (y-axis) of ER+ normal network with ER- normal ceRNA networks built by different cutoff values (x-axis). The bigger the p-value is, the more similar the two networks are. (H) Comparison of the ER+ normal reference network with ER- normal ceRNA networks built by different correlation cutoff values in the average clustering coefficient.

2.4.3 ′UTR Shortening Is Associated With the Aggressive Metastatic Phenotypes of ER- Tumors in ceRNA

In normal ER- ceRNA network based on the subtype-specific co-expression cutoff, 1,783 genes are in the ceRNA relationship with 521 3′US genes (3′US ceRNA partners). Among 1,783 3′US ceRNA partners, 498 (27.9%) are found only in ER- (ER- 3′US ceRNA partners), whereas the other 1,285 (72.1%) are also in ER+ as 3′US ceRNA partners (common 3′US ceRNA partners, **Figure. 2.2A**). We found that 118 IPA canonical pathways significantly ($P < 0.01$) enriched for the ER- 3′US ceRNA partners (**Supplemental Table. 1. Tab 2**) are linked with several aspects of ER- specific tumor phenotypes (**Figure. 2.2B**). The first set of the pathways are "cancer" pathways. For example, the "Molecular Mechanisms of Cancer" pathway ($P=10^{-5.25}$) includes a comprehensive set of genes, disruptions of which are known to promotes tumor growth. Specific to breast cancer, the enrichment of the "Breast Cancer Regulation by Stathmin1" ($P=10^{-3.92}$) pathway is interesting, since overexpression of Stathmin1 correlates with loss of the ER [76] and with aggressive breast tumor phenotypes [77]. The second category of pathways underlies the aggressive metastasis of ER- tumors. For example, among eight pathways that were shown to play

roles in breast tumor metastasis [78], we found that five of them are significantly enriched for ER- 3′US ceRNA partners with the exception of PI3K/AKT, the enriched p-value of which is just below the significance cutoff ($P=10^{-1.95}$). Further, previous studies have associated breast tumor malignancy and poor survival with abnormal control of Ephrin A (reviewed in [79]), which is strongly enriched for ER- 3′US ceRNA partners ($P\text{-val}=10^{-5.05}$). In normal samples without 3′-UTR shortening, 3′US ceRNA partners should closely regulate these pathways. However, in ER- tumors characterized by widespread 3′US events, most (81.7%) of the 3′US ceRNA partners lost the ceRNA relationship (**Figure. 2.2C)**, likely losing the normal control.

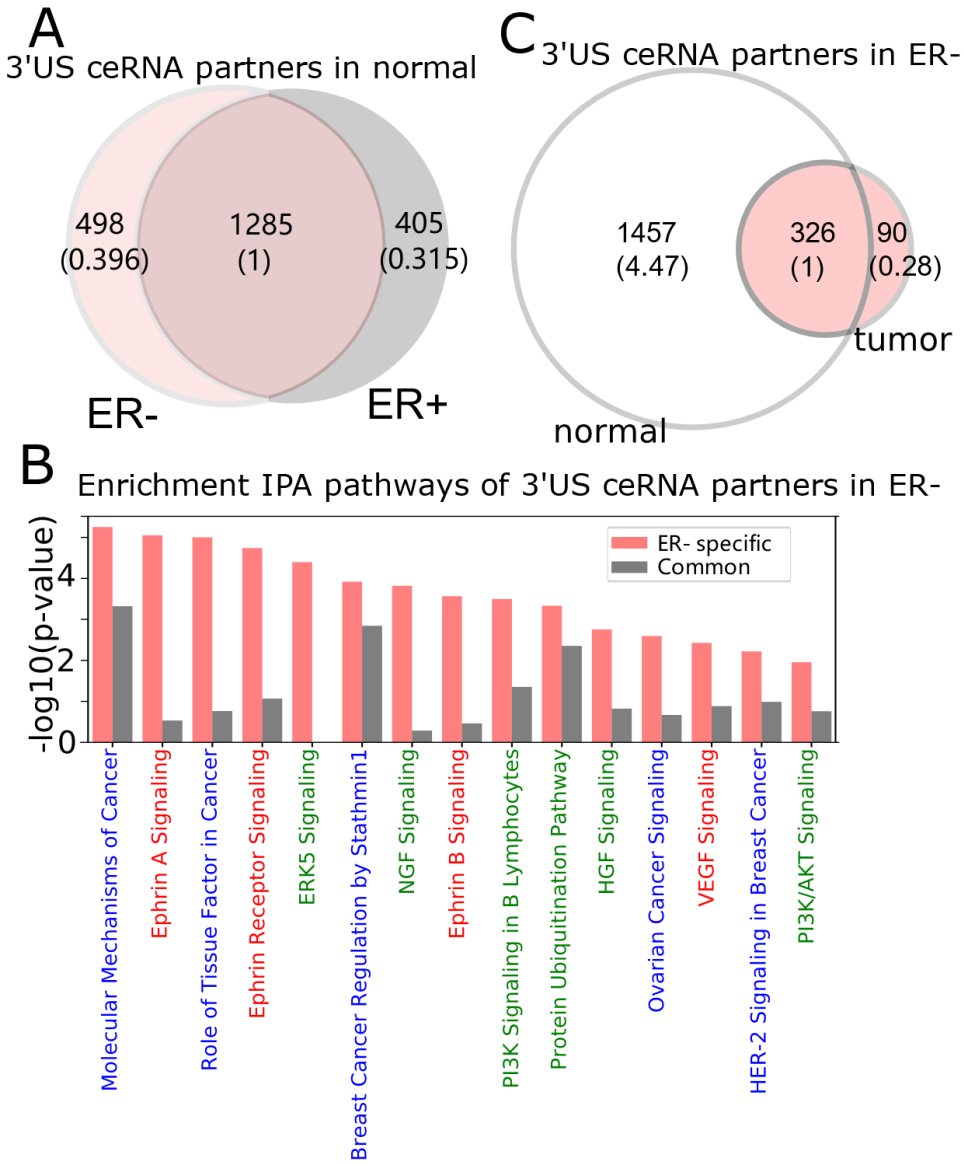


Figure 2.2. 3′UTR shortening is associated to ER-'s aggressive phenotypes in ceRNA. (A) Intersection of 3′US ceRNA partners between ER- and ER+ normal ceRNA networks. (B) IPA canonical pathways significantly (P < 0.01) enriched for the ER- 3′US ceRNAs. Pathways are colorcoded by keyword, "Cancer" in blue, "Signaling" in red and those associated with aggressive phenotypes [78] **in green. (C) Intersection of 3′US ceRNA partners in ER- between normal and tumor ceRNA networks.**

2.4.4 Housekeeping Genes Keep ER+ and ER- Normal ceRNA Networks to Similar Topology

Further, we categorized genes that have possible sponge effect (>5 miRNA binding sites in the 3′UTR) into housekeeping (HK), tumor-associated (tumor suppressors or oncogenes, TA), and transcription factor (TF). Based on 3,804 HK [67], 932 TA [69], and 1,020 TF genes [68] curated in public databases (see Methods), the ceRNA networks consist of 3-fold more HK genes than TA or TF genes (**Figure. 2.3A** for normal and **Supplemental Figure. 4A** for tumor). Due to their active roles in cell maintenance [67], HK genes are expected to maintain constant expression levels under most physiological conditions [67]. Accordingly, the 958 HK ceRNA genes in ER- normal (**Figure. 2.3.A**) express as highly as (**Supplemental Figure. 4B)**, but with less significant variation ($P=1.72e^{-54}$) across the normal samples (**Figure. 2.3B**), than 1,906 non-HK ceRNA genes in the network. With our observation that the HK genes contain more miRNA binding sites than the other genes (P=0.05, **Figure. 2.4C**), they should function as stable sponges for miRNAs [80]. Thus, with a significant number ($P=8.77e^{-771}$) of overlap in the HK ceRNA genes between ER- and ER+ normal samples (**Figure. 2.3D**), we hypothesize that they keep ER- and ER+ normal ceRNA networks in similar topology. To test this hypothesis, we first selected edges involving the HK ceRNA genes from the ER+ and ER- normal ceRNA networks to form subnetworks and compared the subnetworks using normalized Laplacian Matrix Eigenvalue Distribution. Further, we randomly subsampled the same number of edges not involving HK genes 200 times from the ER+ and ER- ceRNA networks and compare the networks in the same way **(Figure. 2.3E)**. The HK ceRNA networks are significantly more similar between ER+ and ER- ($P < 0.01$) than 200 non-HK ceRNA networks, suggesting that HK genes make normal ceRNA crosstalk consistent between the subtypes through the miRNA sponge effect.

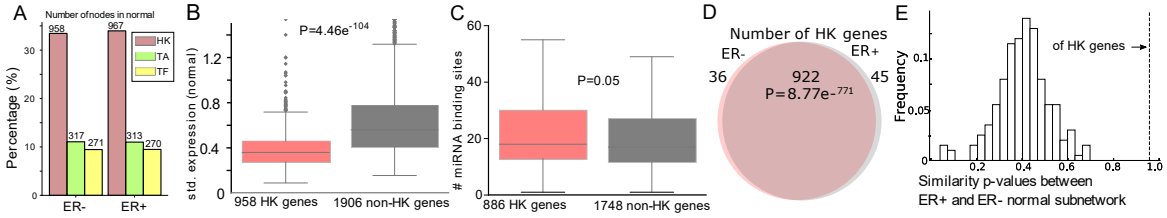
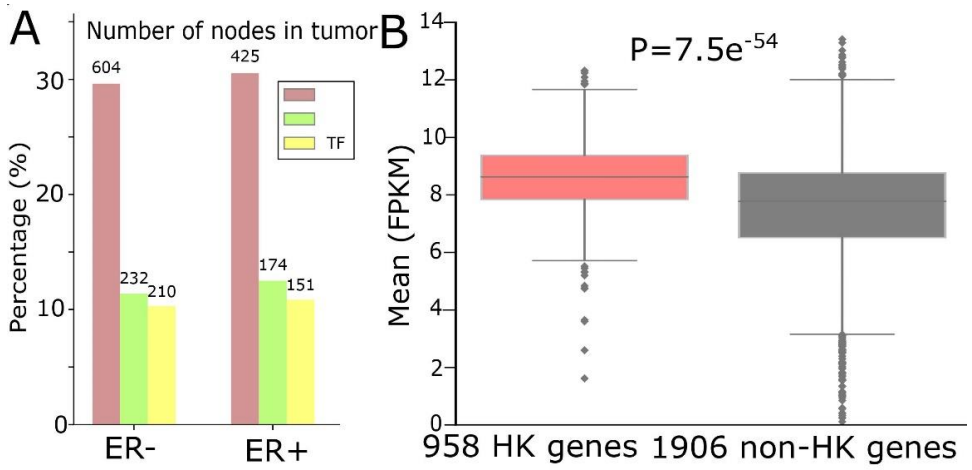


Figure 2.3. Housekeeping genes make consistent ceRNA networks between ER- and ER+ normal samples. (A) Number (and the percentage to the total number of nodes in the networks) of housekeeping (HK), tumor-associated (TA), or transcription factor (TF) genes in the ER- and ER+ normal ceRNA networks. (B) Standard deviation of gene expressions of 958 HK genes and 1,906 non-HK genes in the ER- normal ceRNA network. (C) Number of miRNA binding sites on the 3′UTR of 886 HK and 1,748 non-HK genes in the network. (D) Number of HK genes shared by ER- and ER+ normal ceRNA networks. (E) Distribution of the similarity p-values between the subnetworks of ER+ and ER- normal ceRNA networks with 922 HK genes or the same number of non-HK genes. The higher the p-value is, the more similar the networks are [71]**.**



Supplemental Figure 4. (A) Number (and the percentage to the total number of nodes in tumor networks) of HK genes and other important classes of genes in ER+ and ER- normal ceRNA networks. (B) Average gene expression values of 958 HK genes and 1,906 non-HK genes in the ER+ and ER- normal samples.

2.4.5 3′US Disrupts ceRNA Crosstalk of Housekeeping Genes for ER- Specific Growth

We further examined the impact of 3′US on the role of HK genes. First, 3′US genes are highly connected to HK genes. Out of 958 HK genes, 727 HK genes (75.8%) are connected to 3′US genes, which is in the same scale as the other classes of genes that are known to be regulated by 3′US genes [50], [52] (196 (61.8%) TA genes and 245 (90.2%) TF genes, **Figure. 2.4A**). Also, these HK genes are more highly connected in the network compared to 231 HK genes that are not connected to 3′US genes (**Figure. 2.4B**). Previously, we showed that 3′US represses the ceRNA partners in tumor [52]. Consistently, these HK genes, ceRNA partners of 10.2 3′US genes on average (**Supplemental Table. 1. Tab 5**), are more repressed in tumor than 231 HK genes not connected to 3′US genes (P-value=0.00035, **Figure. 2.4C**). For example, Transforming Growth Factor Beta Regulator 1 (TBRG1) is connected to four 3'US genes (PPP6C, DICER1, H2AFV, UBL3) in ER- normal samples. With 3'US in ER- tumor samples, TBRG1 is significantly down-regulated (logFC = -0.15) considering the general up-regulation of the other housekeeping genes (**Figure. 2.4C**). TBRG1 and those 4 3'US genes are predicted to share binding sites of miR-874 (see **Materials and Methods**). MiR-874 was experimentally shown to repress TBRG1 to promote non-familial breast cancer[81]. Although miR-874 was expressed (avg. FPM is 5.3 and 5.1 in ER- tumor and normal samples), they were not significantly (P-value=0.58) up-regulated in ER- tumor samples to repress TBRG1. Instead, 3'UTR shortening of the four genes may redirect miR-874 to bind more efficiently on TBRG1, leading to its repression. We checked that TBRG1 is not alternatively polyadenylated in ER- tumors (neither 3'US nor 3'UL). Globally, we checked that only 76 out of 958 HK ceRNA genes in ER- (7.9%) are either 3'US or 3'UL genes in tumors. This low overlap between our HK genes and 3'US genes implies that HK genes may not be directly related to growth-related functions [82], [83], but contribute to tumorigenesis through 3'US-

ceRNA. To further understand the impact of the repression on the ceRNA network, we compared the number of the ceRNA partners of these HK genes between normal and tumor. Previously, we showed that 3′US genes will break their relationship with the ceRNA partners [52]. Since the ceRNA relationship changes, either loss or gain, could propagate to neighboring ceRNA relationships [51], the repression of HK genes should break the ceRNA relationship not only with 3′US genes but also with other ceRNA partners. Consistent to the expectation, 727 3′US HK ceRNA partners lost higher ratios of the ceRNA partners in tumor (**Figure. 2.4D**). We found a similar trend of HK gene repression in ER+ breast cancer when connected to 3′US genes (**Supplemental Table. 1. Tab 6**).

The loss of HK ceRNA partners naturally reduces the high overlap of HK genes between ER+ and ER- (**Figure. 2.5A**), resulting into 505 and 144 HK genes that are ceRNA partners of 3′US genes unique in ER- and ER+ tumor (ER- and ER+ HK ceRNA partners), respectively (**Figure. 2.5B**). While it is known that cell growth and cell cycle regulations are different in the subtypes [84]–[86], we found that the 505 ER- HK ceRNA partners are enriched for cell growth- and cell cycle-related IPA pathways (**Figure. 2.5C, Supplemental Table. 1. Tab 3**). First, they are enriched for pathways associated to growth factor (with keyword "GF"). Especially, EGF (P-val=$10^{-2.99}$) activates cell cycle progression in ER- tumors [87], and expression of VEGF (P-val=$10^{-2.42}$) is associated to ER- tumors [88]. Also, both EGF and VEGF are suspected to proliferate ER- tumors when estrogen cannot sustain them [88]. Second, cell cycle pathways are enriched for ER+ specific HK ceRNA partners, suggesting that ER-regulated cell cycle [89], [90] differentiates ER+ and ER- cancer partially at the ceRNA level. Especially, since regulation of cell cycle, G1- and S-phase and their transition ratio, is crucial for ER+ tumor's proliferation (reviewed in [91]), it is interesting that cell cycle regulation pathways for various phases (G1/S or G2/M) of

various mediators (Estrogen or Cyclins) are enriched with 144 ER+ HK ceRNA partners. Third, considering that the enrichment analysis was for the disjoint sets of genes (505 unique to ER- and 144 unique to ER+), it is interesting that these unique HK ceRNA partners are commonly significantly enriched for some "cancer" pathways e.g. "Molecular Mechanisms of Cancer", showing that the HK ceRNAs are involved in cancer mechanisms equally significantly but in a subtype-specific fashion.

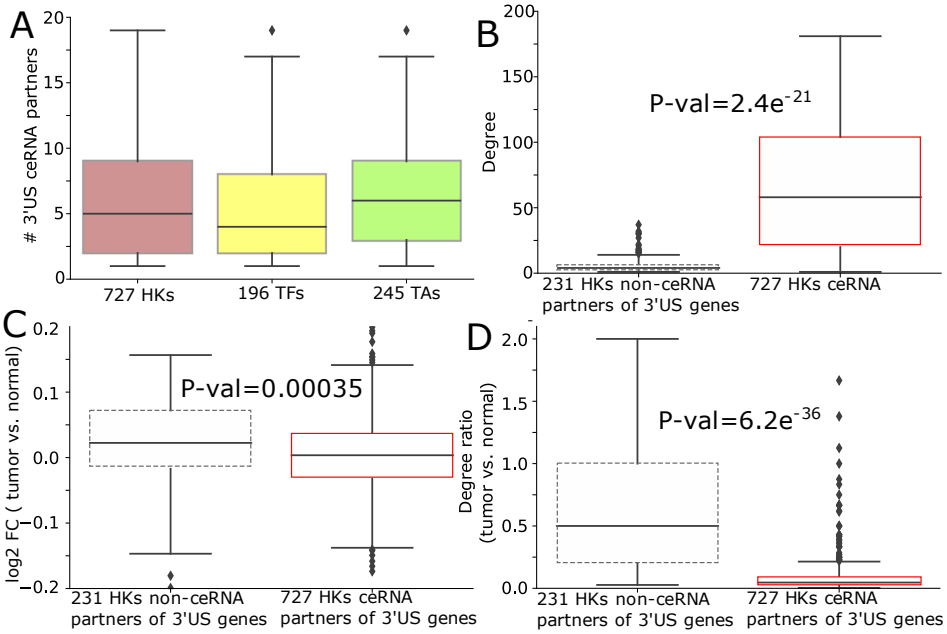


Figure 2.4. 3′US disrupts ceRNA relationship of HK genes in ER- tumors. (A) # of 3′US genes connected to housekeeping (HK), transcription factor (TF), and tumor-associated (TA) genes in the ER- ceRNA network. (B) Degree (# neighbors in ER- normal ceRNA network), (C) log2 fold change (tumor vs. normal), (D) degree ratio (tumor vs. normal) of 727 and 231 HK genes that are ceRNA partners of 3′US genes or not, respectively. Degree ratio in (D) represents the ratio of the number of neighbors retained in tumor.

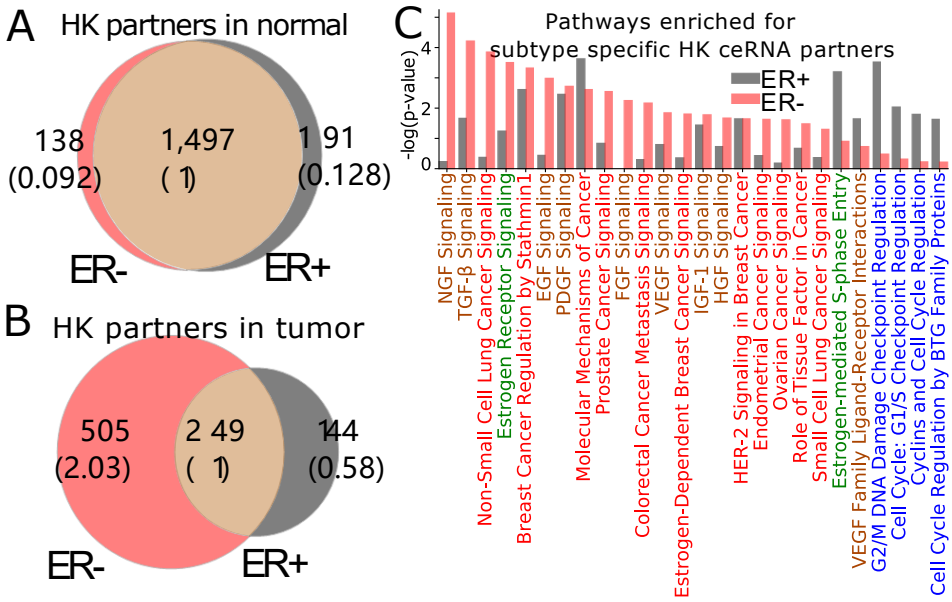


Figure 2.5. 3′US disrupts the ceRNA relationship of HK genes for ER- specific growth. Number of HK ceRNA partners unique and common to ER- and ER+ normal (A) and tumor (B) ceRNA networks. The numbers in parentheses are normalized to the number of genes shared between tumor and normal. (C) IPA canonical pathways significantly (P < 0.01) enriched for ER+ and ER- specific HK ceRNA partners. Pathways are color-coded by keyword, "Cancer" in red, "GF" in brown, "Estrogen" in green, and "Cell Cycle" in blue.

2.4.6 3′US Represses Housekeeping Genes to Promote Tumor Growth

To gain insights into the cause-and-effect relationship from 3′US-mediated HK gene repression to tumorigenesis, we revisited a previous study [52], [82], in which 3′US-ceRNA effect promotes tumorigenesis in NUDT21 knockdown (KD) in HeLa cells and glioblastoma (data available in GSE42420 [14] and GSE78198 [31]). First, we chose 11,431 genes that are expressed in the experiment data (avg. FPKM > 1). Among them, we further chose 4,430 genes that would work as miRNA sponges (>5 miRNA binding sites). To identify ceRNA relationship with the genes, we will solely use significance of miRNA binding site overlap (FDR < 0.05), since the other

criteria for the ceRNA identification, co-expression, cannot be effectively estimated from two replicates of NUDT21 KD experiments. In this way, we identified 860 3′US genes and 2,449 of their ceRNA partners. Among these 3′US ceRNA partners, a significant portion of them (705, 28.8%) are HK genes, while 184 are TA and 163 are TF genes. Especially, it is interesting to note that HK genes in the network are only either 3'US genes (n=298) or 3'US ceRNA partners (n=705). On the other hands, almost half of the TA and TF genes in the network are not connected with 3'US genes (149 of 333 (44.7%) and 147 of 310 (47.4%) for TA and TF, respectively), showing that HK genes can be a major target of 3'US ceRNA effect. Based on our previous finding that 3′US represses the ceRNA partners in tumor [52], we further checked the repression of HK genes in NUDT21 KD. 705 HK genes that are 3′US ceRNA partners are more repressed than TA and TF genes or than 298 HK 3'US genes in the network (**Figure. 2.6A,** P-value=0.01 and 0.05, 0.002, respectively). These results confirm that HK genes are repressed in the tumorigenic process 3′US-ceRNA effect promotes [52].

To assess the impact of this repression on tumor growth, we further conducted IPA analysis on 705 HK 3'US ceRNA partners in comparison to the other 2,410 HK genes not in the network. First, although there are much less HK 3'US ceRNA partners than the other HK genes, they are enriched for more IPA Diseases & Functions terms (**Supplemental Table. 1. Tab 4**). While the IPA analysis gives N/A for the terms that are so lowly enriched that cannot be estimated, HK 3'US ceRNA partners have 581 terms with N/A value and HK genes not in the network have 693 terms with N/A value. Further, we replaced the N/A values with the minimum value and compare the p-values in HK 3'US ceRNA partners vs. the other HK genes. This comparison shows that more terms are significantly (P-value < 0.01) enriched for HK 3'US ceRNA partners (254 terms with better p-values for HK 3'US ceRNA partners and 141 for the other HK genes). This trend is more

pronounced for the terms that are important for cancer. For example, IPA terms with keywords "Cell", "Cancer (or Tumor)", "Apoptosis (, Death, or Necro)", and "Growth (, Proliferation, or Progression)" are significantly (P-value < $2.2e^{-16}$) more enriched in the HK 3'US ceRNA partners, while certain terms for general biological processes such as "RNA" are enriched in the other HK genes (**Figure. 2.6B**). While this analysis does not support our hypothesis as a whole, it demonstrates a potential role of HK gene repression in a tumorigenesis process with HeLa as a model system. It follows that the ER- specific tumor progression is attributable to the repression of different HK genes.

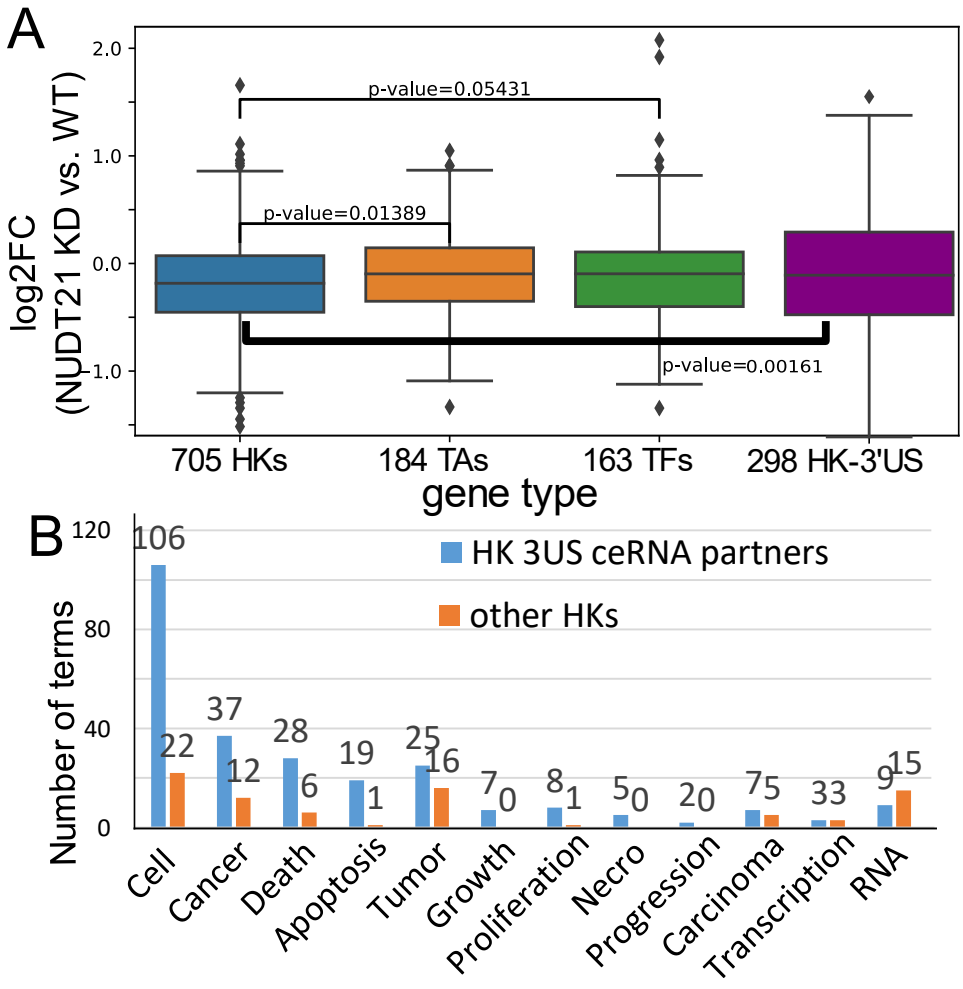


Figure 2.6. (A). log2 fold change (FC) (NUDT21 KD vs. WT) of 705 HK, 184 TA, and 163 TF genes that are (potential) ceRNA partners of 3'US genes. log2FC of 298 HK genes that are 3'US genes is displayed on the

rightmost box. (B). Number of terms with the keyword indicated on the x-axis. Numbers on bar represent the actual number of terms.

2.5 Discussion

To investigate the role of 3′US-ceRNA effect [52] for estrogen receptor negative (ER-) vs. ER+ breast tumors, we built the ceRNA networks that are comparable to each other subtype by addressing the bias due to the different number of samples (72 for ER+ and 20 for ER- in TCGA). A fair comparison of the networks suggests that 3′US disrupts the ceRNA network for ER- tumors' aggressive phenotypes. Further, we revealed a role of 3′US-ceRNA effect on housekeeping (HK) genes. In the cancer context, the potential for being ceRNA was identified for mRNAs (e.g. [92]) as well as long non-coding RNAs (e.g. [93]) and pseudogenes (e.g. [94]). Among mRNAs, it has been shown that tumor-associated (TA) genes and transcription factor (TF) genes heavily contribute to ceRNA regulation [50]. While reaffirming the high contribution (and thus high potential of biological function) of the TA and TF genes to breast cancer ceRNA networks, we further found the high contribution of housekeeping (HK) genes. HK genes were reported as stable "control" genes for miRNA sponge effect (e.g. [95]), indirectly supporting our novel findings. By analyzing TCGA breast cancer and reanalyzing an experimental data, we found more direct supports for their roles to ceRNA.

Further analyses show that 3′US disrupts ceRNA crosstalk of HK genes in a subtype-specific fashion. First, we showed that a subset of HK genes is *trans* target of 3′US-ceRNA effect (sponge HK genes) enriched in important pathways in association to ER-'s aggressive phenotype. Since they are much less than the other HK genes in number (e.g. 705 3′US ceRNA HK genes vs. 2,401 HK genes in the NUDT21 KD experiment), our definition may shed novel insights into identifying another set of biomarkers indicating tumor progression.

In network analysis, a network of interest is often compared to a reference network. However, if the networks are built from different numbers of samples, the comparison will be biased due to the sample size difference (**Supplemental Figure. 3**). With the assumption that normal samples should have similar molecular dynamics, we found the subtype-specific cutoff values for normal ceRNA networks. Then, we construct ER+ and ER- tumor ceRNA networks (two-step pairwise normalization method). As the resulting ceRNA networks facilitate novel discoveries on the subtype-specific 3′US-ceRNA effect, we expect that the two-step pairwise normalization method can further help normalize biological networks built with the different number of samples if the matched normal samples are available.

We note that this normalization method can help further identify the genes playing important roles in a subtype-specific fashion. For example, we used the KS test to compare the eigenvalue distribution of the Laplacian matrix of the two networks, ER+ and ER- ceRNA network. The eigenvalue distribution is a set of eigenvalues each representing a temporal snapshot of the network [71]. Since $K_{1,2}$ in **Eq. 4** represents the snapshot point at which the topology of the two networks is most apart, the edges appearing at that time point strongly differentiate the two subtypes in the ceRNA level. In that sense, genes in the edges can be further investigated for their roles in each of the subtypes. Also, the resulting networks, the comparable ceRNA networks of

ER+ and ER- breast tumors, can further help identify important genes for specific functions in the subtypes. Biological network analysis techniques were used to identify the genes playing important roles in the ceRNA network [50], [96]. To identify such genes for ER- tumor, the samples need to be compared with ER+ in our context. In that sense, we can build a differential network (ER- vs. ER+) based on the comparable ceRNA networks. Then, since hub genes in the differential network would facilitate the ceRNA regulation of many genes only for a specific subtype, e.g. ER- breast tumor, they would be good candidates for important functions specific to the ER- tumors. We can further identify those for specific functions based on the gene sets defined for the functions e.g. Gene Ontology [97]. Our study showed the distinct 3'US-ceRNA dynamics between the ER+ and ER- group of tumor samples. Although ER status is an important clinical variable [39], it is important to note that the two groups do not directly represent further clinical subtypes of breast cancers, such as HER2+ or Triple-Negative. Thus, to reveal further clinical relevance of 3'US-ceRNA dynamics, more study is warranted in the clinical subtypes within each group.

3.0 Project 2 - Deep Neural Networks With Knockoff Features Identify Nonlinear Causal Relations and Estimate Effect Sizes in Complex Biological Systems

3.1 Summary

Learning the causal structure helps identify risk factors, disease mechanisms, and candidate therapeutics for complex diseases [98]–[100]. However, although complex biological systems are characterized by non-linear associations, existing bioinformatic methods of causal inference cannot identify the nonlinear relationships and estimate their effect size [101]–[105]. To overcome these limitations, we developed the first computational method that explicitly learns nonlinear causal relations and estimates the effect size using a deep-neural network approach coupled with the knockoff framework [106], named causal Directed Acyclic Graphs using deep-learning VAriable SElection (DAG-deepVASE) [107]. Using simulation data of diverse scenarios and identifying known and novel causal relations in molecular and clinical data of various diseases, we demonstrated that DAG-deepVASE [107] consistently outperforms existing methods in identifying true and known causal relations. In the analyses, we also illustrate how identifying nonlinear causal relations and estimating their effect size help understand the complex disease pathobiology, which is not possible using other methods. With these advantages, the application of DAG-deepVASE [107] can help identify driver genes and therapeutic agents in biomedical studies and clinical trials.

3.2 Background

The most important thing to consider when adding elements to your ETD, is to aim for consistency. If you add block text or quotations that vary from the Normal style, your best bet is to create a style for that customization and use it throughout the document. It's also best to minimize the amount of in-line editing that you do, as when you adjust a few lines of text that varies from the rest of the document, it will most probably be flagged on review.

Since molecular and clinical variables interact for the development of complex diseases such as cancer, asthma, and sepsis [98]–[100], learning the causal structure among the variables helps identify risk factors, disease mechanisms, and candidate therapeutics for the complex diseases for future evaluation. For example, if an abnormal expression of a certain gene modifies the expression level of other genes and contributes to the development of a disease, then controlling this gene can lead to the effective treatment of the disease.

A popular statistical model for causal inference is the causal directed acyclic graph (DAG), which learns conditional dependence among variables[105], [108], [109] because the conditional dependence can further imply the causal relationships under three causal assumptions: Markov, faithfulness, and sufficiency. The causal Markov condition states that causal relationships among the set of variables in their probability distributions (e.g. Bayesian network) are conditionally independent of their non-descendants given their parents[110]. The causal faithfulness condition states that all independence relations in the data are consequences of the Causal Markov condition. The causal sufficiency condition states that input data measured all the common causes of the

measured variables, thus no latent (unobserved) confounder exists. Since the assumptions are not usually met in data, statistical causal inference is limited to identifying causal relationships that are Markov equivalent, which hold the same adjacencies and imply the same independence and conditional independence relationships on the same variables (v-structure). Under the assumptions, bioinformatic methods have incorporated two main approaches to building DAGs: constraint-based or score-based[27], [36], [111]–[116]. Constraint-based algorithms learn constraints that restrict the set of possible causal graphs by testing conditional independence in the input data. Peter and Clark (PC) [27], one of the most popular algorithms under this category[28]–[33], uses a combination of conditional independence tests and graph pruning techniques first to determine an skeleton of the DAG and then to determine the causal directions in the skeleton network. On the other hand, score-based algorithms generally formulate the causal learning problem as a search problem to optimize a certain score function with respect to an unknown DAG and the input data. For example, the Degenerate Gaussian score (DG) was recently proposed [34] by extending the widely used BIC score [35], [36] for mixed types of data. Specifically, by embedding discrete variables into a continuous space using one hot vector representations, DG demonstrates a near-perfect performance under certain simulation scenarios of high-dimensional data.

Previously, causal inference methods have been successfully used to provide insights into molecular mechanisms and predict treatment effects. First, to provide insights into molecular mechanisms (e.g., transcriptional regulatory relationship between genes), methods have been developed to integrate multiple types of data where the direction of effect is known from one type to another (e.g., from DNA variants to gene expression). Developments in this approach utilized both score-based [117], [118] and constraint-based [119]–[122] algorithms. For example, MRPC

uses PC to examine a set of causal relationships between DNA variants and gene expression information implied by the principle of mendelian randomization (PMR). Second, to predict treatment effects, causal inference was done using multiple intervention trial or experiment data (e.g., RNAi-based gene knockout experiments). For example, conservative local causal discovery (CLCD) tests conditional (in)dependence among multiple entities (e.g., proteins) across experiments [123] and BACKSHIFT evaluates particular causal scenarios shared across experiments using a linear causal model [124]. Treatment effect can also be predicted based on the relationship of the input samples with other samples for which treatment effects are known. To this end, causal k-nearest neighbor algorithm (causal KNN) estimates the effect based on the nearest neighbors with known treatment effects. Similarly, causal random forest attempts to identify neighbors after recursively partitioning the covariate space through creating a set of decision trees. While they successfully identified biologically meaningful or clinically reasonable causal relations in various validation experiments, they are not necessarily relying on artificial intelligence (AI) methods and tend to test independence for a limited number of entities or based on naïve assumptions on the relationships among data points.

Recently, methods have employed AI methods to address the limitations of previous causal inference methods. For example, a recent development, causalMGM (causal mixed graphical model) [125], first identifies associations between different types of data using a mixed graphical model (MGM) and then infers causality of the associations through PC. This two-stage approach showed good scalability and accuracy for high-dimensional simulated and biological data of mixed types [125]. Also, to identify an optimal DAG based on an optimality score, a challenge can be the intractable search space that increases with a complexity super exponential to the number of the input variables. Thus, a group of methods have been developed to efficiently navigate the search

space. Previously, this problem was addressed with additional structure assumptions, e.g., in terms of tree width [126], number of variables [127], ancestral constraints [128], or a set of prior knowledge [129]. While they were designed to shrink the intractable search space with the assumptions, methods can also be developed to expand the search space and efficiently navigate it. In that regard, a recent breakthrough formulates the problem as a continuous optimization with a structural constraint that ensures acyclicity [130] and spurs further development of deep neural network (DNN) models. For example, Yu et al. proposed a deep generative model and apply a variant of the structural constraint to learn the DAG (DAG-GNN) [37] and Zheng et al. generalized this framework so various approximations can be used for search (NOTEARS) [38] , including neural networks. Despite all substantial progresses in those approaches, we found several challenges to identify causality for complex diseases. First, a method should identify both linear and non-linear associations. While linear associations may exist, complex biological systems are characterized by non-linear associations [21], [22]. For example, the effects of hormone receptor status on breast cancer biology are often nonlinear due to their complex interactions with other molecular complexes in multiple regulation processes[23]–[25]. Some of the nonlinear associations may be revealed in the existing DNN methods. However, the methods utilize the DNN component to effectively navigate the search space over various DAGs while optimizing an optimality score across all the relationships in a DAG as has been done previously. Generally, the optimality scores are based on a likelihood model with product terms to represent the variable relationships. For example, as the optimality score, both DAG-GNN and NOTEARS can use the BIC score that select the product terms to determine significant variable relationships. A product term of two variables assumes that the relationship between the variables is additive and proportional, meaning that the effect of one variable on the other is assumed to be constant across

all levels of the other variable. Since the constant effect is satisfied only in linear relationships, the methods based on such optimality scores are designed to consider only linear relationships. In other words, as a DNN component is to address nonlinearity, existing methods use the DNN component to address nonlinearity in how various DAGs are searched through, not to address nonlinearity in each relationship. In this sense, they do not explicitly identify each causal relationship as nonlinear. Second, a method should estimate the effect size of each association. This is critical to facilitating a translatable understanding of the causal relationships since it is important to select a limited number of the most significant causal relationships for downstream experiments or clinical trials due to both technical and practical limitations. However, currently, no method can not only identify the nonlinear relationships but also estimate their effect size.

To address these limitations and enable a more realistic and translatable causal structure learning for complex diseases, we developed the first computational method that explicitly learns nonlinear causal relationships as well as linear causal relationships, named causal Directed Acyclic Graphs using deep-learning VAriable SElection (DAG-deepVASE) [107]. To identify nonlinear causal relationships in high-dimensional data, DAG-deepVASE [107] incorporated a two-step approach: 1) identify associations and estimate their effect sizes and 2) infer the causality among the associations. In the first step, to identify each causal relationship as nonlinear, DAG-deepVASE [107] puts a deep neural network (DNN) model between each potential causal relationship. However, a regular DNN model cannot estimate the effect size between an input variable and the response variable since it would be difficult to summarize the edge weights between neurons across multiple layers between the variables. To address this difficulty, DAG-deepVASE [107] incorporated the knockoff framework into the DNN model to estimate the effect size. Previously, this architecture was used to control false positive rate in the context of variable

selection [131]. In this work, we extend this architecture to measure the effect size in the context of causal inference for the first time. Further, to learn the causal direction for the identified nonlinear associations, DAG-deepVASE [107] extends a score-based approach, DG. While it was not known which causal inference approach would learn the causal direction of nonlinear associations, we conducted extensive studies to find that its asymptotic properties make the inference tractable and flexible enough to learn nonlinear causalities.

DAG-deepVASE [107] consistently outperforms other methods in identifying true causal relations in simulation data of diverse scenarios and identifying known and novel causal relations in molecular and clinical data of various diseases (pediatric sepsis, gut bacteria/nutrient intake and BMI, and breast cancer), facilitating a systematic understanding of the complex disease pathobiology. In the analyses, we also illustrate how identifying nonlinear causal relations and estimating their effect size help understand the complex disease pathobiology, which is not possible using other methods.

3.3 Materials and Methods

In developing our method, we followed the DOME (Data, Optimization, Model, and Evaluation) guidelines stated in https://dome-ml.org/. Especially, we selected testing data that is representative of the domain (TCGA breast cancer for molecular data, gut microbiome and obesity data for metagenomics, and pediatric sepsis data for clinical data) per the Data guidelines. Their accessions are further detailed in the **Availability of Supporting Data** section below. Per the Optimization guidelines, we performed experiments with various numbers of neuron layers (1~5 layers) and various numbers of neurons (10, 50, 100, 200, 400, and 600 neurons) in each layer on

the simulation data (10 and 190 features in true and false causal relation to the outcome, respectively, generated for 1,000 samples) (**Supplemental Figure. 8**). These experiments justify the current design principle of DNN methods to put multiple layers of the neurons that is the same as the number of input features. For example, our experiments demonstrate that, to run on the simulation data, which consist of 100 features, DNN models of multiple layers of 100 neurons perform the best. In this project, we followed this design principle to implement our methods and reported all hyperparameters (**Table 3.1**) and optimization protocol under **Running parameters of DAG-deepVASE** [107] section below. Per the Model guidelines, we dockerized our method to make it easier for people to test and deploy. Lastly, per the Evaluation guidelines, we compare our method both with public method (causalMGM) and simple (baseline) method (linear-DG) on the same dataset.

Table 3-1. Parameter settings for the deep-learning component of DAG-deepVASE.

	Parameters	Value
DNN	Activation function	Rectified linear unit (ReLU)
	Initial weight values	Glorot normal intializer
	Regularization	$L1{-}regularization$
	Optimization	Adam optimization
	Loss function	Mean of squares of errors (MSE)
FDR	FDR control rate	0.05

3.3.1 Availability of Supporting Data

We used a simulation data set, two public data, and one access-controlled data. Our simulation data are downloadable from our project website (https://github.com/ZhenjiangFan/DAG-deepVASE). TCGA breast invasive carcinoma (BRCA) data were downloaded from https://tcga.xenahubs.net, available under BRCA cohort, under gene expression RNAseq section, on IlluminaHiSeq (n=1,218) TCGA Hub. It consists of the gene expression RNAseq dataset (dataset ID: TCGA.BRCA.sampleMap/HiSeqV2) and the clinical phenotype dataset (dataset ID: TCGA.BRCA.sampleMap/BRCA_clinicalMatrix). To investigate the dietary effect of the human gut microbiome, we downloaded a cross-sectional data of 98 healthy volunteers from https://noble.gs.washington.edu/proj/DeepPINK/ that preprocessed the data set collected from [132] . We also used an access-controlled data of pediatric sepsis. The entire data are available upon request and after taking due steps for the rights and welfare of human research subjects involved in the study (regarding the Institutional Review Board review). However, to ensure reproducibility of our findings, we uploaded a down-sampled (70%) version of our data sets for the interactions of SIRS on the code and data repository site described below. The interactions of SIRS forms the basis of our novel findings and include SIRS with heart rate, CRP (C-reactive protein), IFN-γ (interferon gamma), CNS (central nervous system) dysfunction, and IL (interleukin)-22. We ensured that our findings are reproduced using this data set. Details of each data are given below.

3.3.2 Breast Cancer Data

The gene expression RNAseq section in the Xena website is the level 3 data estimates in log2(x+1) transformed RSEM normalized count obtained from the TCGA data coordination centers. The University of North Carolina TCGA genome characterization center experimentally measured the gene expression profile using the Illumina HiSeq 2000 RNA Sequencing platform. Since we selected genes based on the expression variation, we did not use gene expression data with further normalization in the Xena website, such as pancan normalization or percentile normalization. For the gene expression dataset, we selected 500 or 2,000 expressed genes based on their variances. Then, we added ERBB2 (also known as HER2 or *neu*) to the selected gene set to the 500 genes selected above. ERBB2 was included due to its important role in human malignancies, especially for human breast cancers [133]. For the clinical dataset, we used 10 well-known clinical status features: PAM50 status (PAM50Call_RNAseq), HER2 status (HER2_Final_Status_nature2012), tumor stage (Converted_Stage_nature2012), tumor node status (Node_nature2012), the progesterone receptor status (breast_carcinoma_progesterone_receptor_status), the estrogen receptor status (breast_carcinoma_estrogen_receptor_status), the number of lymph nodes (lymph_node_examined_count), neoplasm cancer status (person_neoplasm_cancer_status), pathologic stage information (pathologic_stage).

3.3.3 Gut Microbiome Data

This data has 214 micronutrients and 87 genera from 90 healthy donors. They were between the ages of 18 and 40 and required to be free from any chronic gastrointestinal disease,

cardiac disease, diabetes mellitus or immunodeficiency diseases, to have a normal bowel frequency (between once every 2 days and 3 times per day), to have body mass index (BMI) between 18.5 and 35. They had not taken antibiotics within 6 months prior to enrollment, proton pump inhibitors, H2 receptor antagonists, tricyclic antidepressants, narcotics, anticholinergic medications, laxatives, or antidiarrhea medications within 4 weeks of enrollment, or NSAIDs, dietary supplements, or antacids within 2 weeks prior to enrollment.

The BMI data were evaluated based on the donors' information and the bacteria data are extracted using 16S rRNA sequencing from the stool samples. For a consistent result with previous analyses, we used the same data pre-processing procedure as previous computational work on the data [131]. In particular, the nutrient values are normalized using the residual method to adjust for caloric intake and then standardized [134]. Then, this data is log-ratio transformed to get rid of the sum-to-one constraint and then centralized. Following [135], 0s are replaced with 0.5 before converting the data to a compositional form. With both the nutrient intake and genera composition as predictors, we treat BMI as the response.

3.3.4 Pediatric Sepsis Data

The pediatric sepsis data were collected from 9 PICUs in the Eunice Kennedy Shriver National Institutes of Child Health and Human Development Collaborative Pediatric Critical Care Research Network (including Children's Hospital of Pittsburgh, Children's Hospital of Philadelphia, Children's National Medical Center, Children's Hospital of Michigan, Nationwide Children's Hospital, Children's Hospital of Los Angeles, St. Louis Children's Hospital, C. S. Mott Children's Hospital, and Mattel Children's Hospital at the University of California Los Angeles) [136]. Briefly, we collected blood samples and clinical data obtained from our previously

published PHENOMS study[136]. Approval was obtained from The University of Utah Institutional Review Board, Central IRB # 70976. Written informed consent was obtained from one or more parents/guardians for each child. Assent was garnered when the child was able. Patients were enrolled from 2015 to 2017. The CONSORT diagram and details of the clinical study protocol have been previously published [136]. In brief, children qualified for enrollment in PHENOMS if they 1) were between the ages of 44 weeks gestation to 18 years of age; 2) were suspected of having infection meeting two or more of four systemic inflammatory response criteria [137], and 3) had one or more organ failures [138]. Three consented and enrolled children who were excluded from reporting in the parent study manuscript due to a maximum per site enrollment of 81 patients to evenly distribute enrollment among the centers, are additionally included in this analysis. Another work investigating this data is in progress and thus this data is currently not deposited in the public domain yet. There originally were 55 candidate clinical features and 33 cytokine features measured from 404 children admitted. We removed features with a missing rate higher than 20% as well as highly correlated features (Pearson's correlation coefficient > 0.6). Finally, we dropped samples with any missing data. As a result, this dataset provides 56 features (**Table 3.2)** from 281 samples with low correlations (< 0.3 and > -0.44 in Pearson's correlation coefficient). In our analyses, some clinical terms were reported with abbreviations; GCS: Glasgow Coma Scale; CRPH: C-reactive protein; SIRS: Systemic Inflammatory Response Syndrome: sCD163: soluble CD163; M-CSF: Macrophage colony-stimulating factor. **Table 3.2** has all the variables with full names. To address the high right-skewness of the clinical data, we employed the log transformation (log10) on the values.

Table 3-2. Variables in the pediatric sepsis data

Variable	Description of variable

Demographic	
Age	
PRISM[a]	
Low SBP	Lowest Systolic Blood Pressure
High Heart Rate	Highest Heart Rate
Low Temp	Lowest Temperature
High Temp	Highest Temperature
GCS	The lowest GCS score
Lower Platelet	Lowest Platelets
Labs	
Higher Creatinine	Highest value from PRISM High Creatinine and High Creatinine
Low Lymphocyte	Absolute lymphocyte count
Low Hemoglobin	Hemoglobin
Low Platelet	Platelet count
ex vivo TNF-α	Blood endotoxin-stimulated TNF-α
SFASLigand	sFas Ligand

sCD163	Soluble CD163
ADAMTS13	A disintegrin and metalloproteinase with a thrombospondin type 1 motif, member 13
Organ failure	
SIRS	Systemic Inflammatory Response Syndrome criteria
Cytokine	
CRP	C-reactive protein
IFN-β	Interferon-β
IL-22	Interleukin-22
IL-18	Interleukin-18
IL-18BP	Interleukin-18-binding protein
MIG-CXCL9	Chemokine (C-X-C motif) ligand 9 (CXCL9) or monokine induced by interferon gamma (MIG)
IL-1β	Interleukin 1β
IL-4	Interleukin-4
IL-6	Interleukin-6
IL-8	Interleukin-8

IL-10	Interleukin-10
IL-13	Interleukin-13
IL-17A	Interleukin-17A
IFN-γ	Interferon-γ
IP-10/CXCL10	C–X–C motif chemokine 10 (CXCL10) or interferon γ-induced protein 10 kDa (IP-10)
MCP-1/CCL2	Chemokine (C-C motif) ligand 2 (CCL2) or monocyte chemoattractant protein 1 (MCP1)
MIP-1α	Macrophage inflammatory protein-1 alpha
MIP-1β	Macrophage inflammatory protein-1β
TNF-α	Tumor necrosis factor α
MCP-3	Monocyte chemotactic protein-3
IFN.α2	Interferon α-2
IL-1α	Interleukin 1α
IL-2Ra	Interleukin-2 receptor antagonists
IL-3	Interleukin-3
IL-16	Interleukin-16

M-CSF	Macrophage colony-stimulating factor
SCF	Stem cell factor
Trail	Trial
Ferritin	Ferritin

3.3.5 Availability of Supporting Source Code and Requirements

Project name: DAG-deepVASE

Project home page: https://github.com/ZhenjiangFan/DAG-deepVASE

Operating system(s): Platform independent

Programming language: Python, Java, C, and R

Other requirements: e.g., Java 1.3.1 or higher, Tomcat 4.0 or higher

License: MIT license

3.3.6 Pre- and Post-processing

To reduce false-positive discoveries, DAG-deepVASE carries out several pre- and post-processing steps. As a pre-processing step, DAG-deepVASE filters out variable pairs that are conditionally independent on all the other variables based on inverse covariance (< 0.0001), using a python function in the package for machine-learning optimization (scipy.linalg.inv). Although it's a common practice for computational causal inference under certain assumptions and the filtered-out nodes may not change the rest of the network, we made it optional since they can be

important nodes for downstream analysis [139]. As another optional postprocessing step, DAG-deepVASE can detect a cycle (a non-empty tail in which the first and the last nodes are equal) in the network connecting the causal relations. Further, users can remove the cycle components by removing edges with their prior knowledge or DAG-deepVASE can automatically remove the edges with the least effect size it estimates.

3.3.7 Directed Acyclic Graph Using Deep-Learning-Based Variable Selection (DAG-deepVASE)

We provide here a brief overview of DAG-deepVASE that aims to identify linearly and nonlinearly associated variables while estimating their effect sizes (**Figure 3.1B, C, respectively**) and learn their causal directions (**Figure 3.1D**) to produce a DAG from data matrix X consisting of M input variables (**Figure 3.1A**). In the first step, to identify linearly associated variables, DAG-deepVASE develops a penalized regression function with the interaction terms connecting the variables and maximizes the likelihood score with sparsity penalties (**Materials and Methods**, **Figure 3.1B**). While the linear associations have been the main focus of previous causal inference methods[102], DAG-deepVASE further identifies nonlinearly associated variables by developing a set of deep neural network (DNN) models, each with one of the input variables as the outcome and all the others as the dependent variables of the model (**Figure 3.1B)**. Note that this approach is different from most existing DNN-based causal inference methods in that DAG-deepVASE models nonlinearity in individual variable relationships while other methods model nonlinearity in the way variable relationships are combined with respect to the input data. Further, we set out to estimate the effect size on the individual variable relationships in our DNN model. Although estimating the effect size is important to design further clinical trials and/or experimental

validations with strong drivers, it is not straightforward to summarize the edge weights across multiple layers for effect size estimation in a regular DNN approach. DAG-deepVASE successfully estimates the effect size of the nonlinear associations by embedding the knockoff variables in the DNN model (**Figure 3.1B**). Knockoff variables are a synthetic and noisy copy of the input variables, which resemble the correlation structure of the input variables, but are conditionally independent of the outcome, given the input variables. This property of knockoff variables allows us to estimate how important the original association is in reference to the knockoff variables, leading to the effect size estimation.

In the second step, after identifying both linear and non-linear associations, DAG-deepVASE determines their causal direction using a single metric to ensure causal inference consistency between linear and nonlinear causalities. Since this is one of the first methods that identify nonlinear causal directions, it is unknown whether PC or DG would work better to identify nonlinear causal directions. Among various measures, we chose to use DG because it is accurate, decomposable, and flexible. While its accuracy, which was demonstrated in simulations[34], is clearly beneficial to learning accurate causal directions, we separately conducted an extensive study to find that its decomposability and flexibility were critical to identifying nonlinear causal directions. DG decomposes the task of identifying the optimal causal structure into determining the causal direction of each association. Whereas PC determines the optimal causal structure by considering all associations simultaneously, decomposability allows us to determine the causal direction of each nonlinear association without referring to other associations, making each causal inference tractable. DG also shows flexibility in learning the causal structure generated outside of its model class (conditional Gaussian model). This flexibility allows us to extend DG to learn nonlinear causal directions. In simulation data of diverse scenarios and biological data of various

contexts, we demonstrate that DAG-deepVASE can learn causal relations up to the Markov equivalence classes of the true causal relationship.

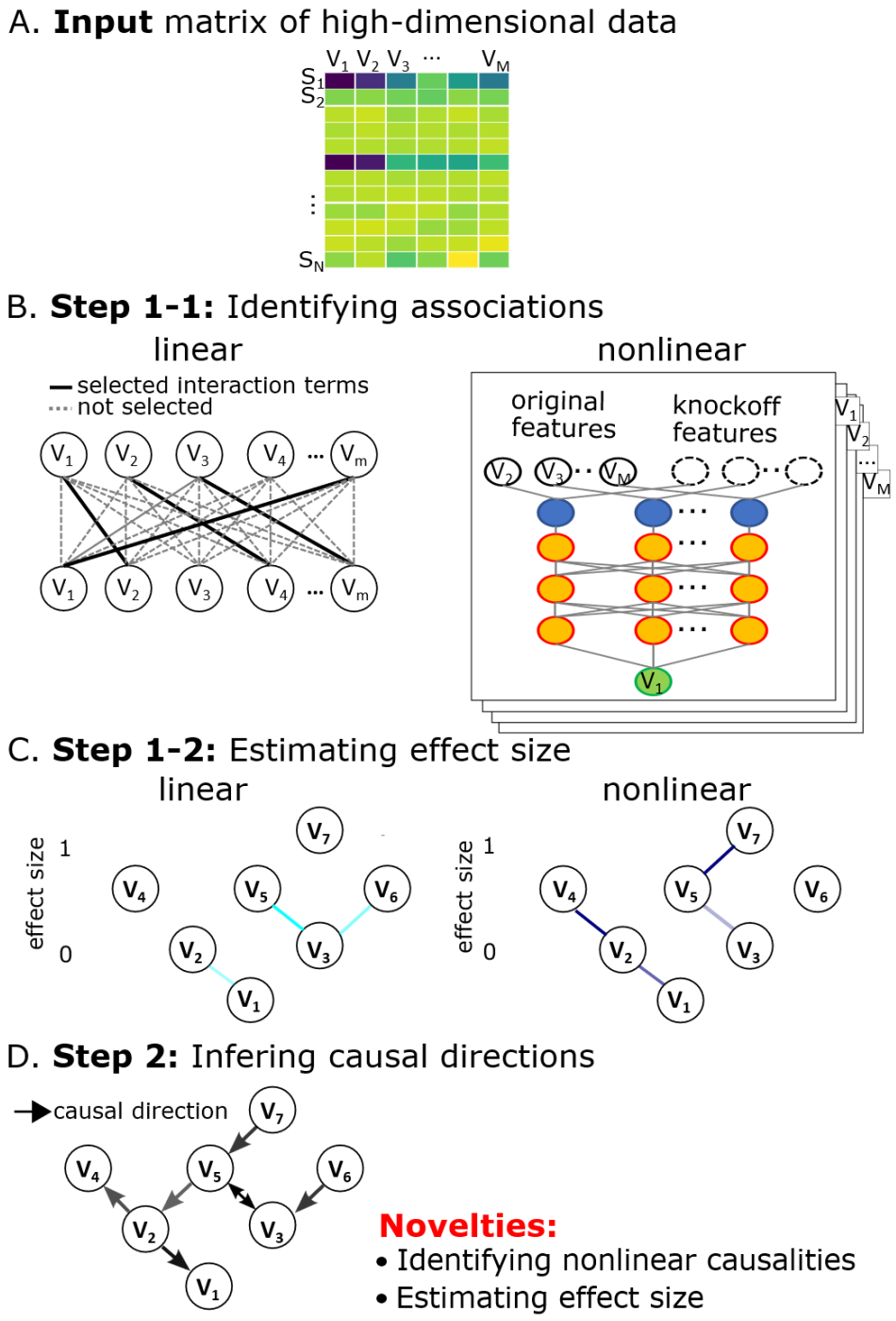


Figure 3.1. Overview of DAG-deepVASE. (A) An input data matrix consisting of *M* variables (V_1, V_2, ..., V_M), either continuous or ordinal categorical, collected from *N* samples. (B) Left: An example of the identified

linear associations using a statistical graphical model (MGM). Right: identifying nonlinear associations using a deep neural network (deep-learning) model. After the first run sets V_1 as response and identifies its association with other variables, DAG-deepVASE will run this model with each of the other variables (V_2, V_3, ..., V_M) as response and with all the other variables as input. (C) Left: estimating the effect size of linear associations in the statistical graphical model. Right: estimating the effect size of nonlinear associations in reference to knockoff filter implemented in the deep-learning model. (D) Learning the causalities by running the degenerate Gaussian (DG) separately on the identified associations, either linear or nonlinear.

3.3.8 Running Parameters of DAG-deepVASE

To identify linear and nonlinear associations in each data set, we first performed pre-processing steps described in the "**Pre- and post-processing**" section below. To identify linear associations after the steps, we ran Lee and Hastie's log-likelihood model [103] for all possible variable pairs (x, y in **Equation 2.1**) with the penalty to select important variables (**Equation 2.2**). We set the sparsity penalty values of the likelihood function to 0.3 unless specified otherwise. Variable pairs remained after applying the penalty are significant linear associations.

To identify nonlinear associations, we first built a deep neural network model consisting of the input layer, two hidden layers, and the output layer (Step 1-1 in **Figure. 3.1)**. Assume that the input data has *p* input variables, then we set the input layer with 2**p* neurons, since we generated the knockoff variable for each input variable and combine them in a pair-wise fashion in the input layer (**Equation 2.3, 2.4)**. The combined input-knockoff neurons are fully connected to the hidden layers. For the case of *p* input variables, each hidden layer has *p* neurons, further transformed using the rectified linear unit activation (ReLU) function[140]. The initial weights for the hidden layer are generated using the Glorot normal initializer[141], which uses

$L1{-}regularization$ with the regularization parameter set to $O(\sqrt{\frac{2logp}{n}})$. To train this model, mean of squares of errors (MSE) is used to calculate the loss in comparison with the response on the output layer. To train the model's parameters with respect to the loss function, we used a stochastic gradient descent method called "Adam optimization". Then, we ran it to identify variables that predict the outcome variable with a high effect size estimated in **Equation 2.5, 2.6.** Equations are described in the section of **Algorithm of DAG-deepVASE**. All running parameters for the nonlinear module is summarized in **Table 3.1**. To identify nonlinear associations between all pairs, we ran this procedure repeatedly with each variable as outcome and all the rest as input. While the procedure was previously developed to identify input variables that can predict the outcome [131], DAG-deepVASE identify these prediction pairs as associated variables based on a widely accepted notion that a predictor and the outcome is statistically an associated pair. For theoretical understanding of the Equations, readers are referred to the following section.

3.3.9 Algorithm of DAG-deepVASE

Let X be the data matrix of interest with variables measured over N observations. $x_i \in X$ is the M -dimensional feature vector observed for sample i, consisting of C continuous variable set X_C and D ordinal categorical variable set X_D $(C + D = M)$. To systematically construct a DAG from both linear and nonlinear associations among variables, DAG-deepVASE leverages a well-established computational framework where variable associations are first identified, and their causal directions are then learned [102], [104], [142].

In the first step, DAG-deepVASE selects linearly associated variables based on Lee and Hastie's log-likelihood [103] as follows.

Equation 2.1. $logp(X_C, X_D, \Theta) = \sum_k^C \sum_l^C \left(-\frac{1}{2}\beta_{kl} X_{Ck} X_{Cl}\right) + \sum_k^C \alpha_k X_{Ck} + \sum_k^C \sum_l^D v_{kl}(X_{Dl}) X_{Ck} + \sum_k^D \sum_l^D \Phi_{\mathrm{kl}}(X_{Dk}, X_{Dl}) - \log(Z)$,

where Θ represents all of the model parameters, β_{kl} is the interaction coefficient between two continuous variables, X_{Ck} and X_{Cl}, α_k is the potential of continuous variable X_{Ck}, v_{kl} is the interaction parameter between continuous variable X_{Ck} with each index of the categorical variable X_{Dl}, Φ_{kl} is a matrix of interaction parameters between discrete variable X_{Dk} and X_{Dl} (indexed by their levels) [103]. If the data consists only of continuous variables, this model reduces to a multivariate Gaussian model with β_{kl} coefficient as entries in the precision matrix. If only with categorical variables, this model is the popular pairwise Markov random field with potentials given Φ_{kl}. While calculating the partition function Z can be expensive, it is possible to optimize the log-likelihood edge by edge [105] under the faithfulness and causal Markov assumptions. Overall, this equation models the log-likelihood of interactions of continuous variables and categorical variables as a multinomial linear regression. To ensure sparsity and select associated variables in the regression model, Sedgewick et al. introduced sparsity penalties for associations between continuous variables, between a continuous and a categorical variable, and between categorical variables ($\lambda_{cc}, \lambda_{cd}, \lambda_{dd}$, respectively) as follows [105].

Equation 2.2. $minimize_{\Theta} \tilde{l}(\Theta) + \lambda_{cc} \sum_{i<j} |\beta_{ij}| + \lambda_{cd} \sum_{i,j} \left|\left|v_{ij}\right|\right|_2 + \lambda_{dd} \sum_{i<j} \left|\left|\Phi_{\mathrm{ij}}\right|\right|_F$

For balance estimation of the associations, DAG-deepVASE uses the same sparsity penalty (0.3 for all three interactions) and set FDR level q as 0.05. After selecting the interactions, we report as effect size the coefficients in the model (β_{ij}, v_{ij}, or, Φ_{ij}, corresponding to the type of the selected variables).

In the second step, DAG-deepVASE selects non-linearly associated variables as follows. To identify nonlinear associations with x_i, DAG-deepVASE sets multiple perceptron layers

between $X_{\setminus i} = \{x_1, x_2, \dots, x_{i-1}, x_{i+1}, .., x_M\}$ and x_i (Step 1. Nonlinear association in **Figure 4.1**) and estimate the effect size of the association between ($x_j \in X_{\setminus i}, x_i$). To estimate the effect size, DAG-deepVASE generates model-X knockoff[143]. For input variables x_i and x_j , the exchangeability property ensures that $(x_i, x_j, \tilde{x_j}) =^d (x_i, \tilde{x_j}, x_j)$, where " $=^d$ " denotes equality in distribution. This exchangeability properties help prioritize causal relations with x_i over simple correlations. For example, suppose (x_i, x_k) is a correlation without causal relation. Then, the feature exchangeability $(x_i, x_k, \widetilde{x_k}) =^d (x_i, \widetilde{x_k}, x_k)$ will hold and make their relationship measure $|RI_{ik}|$ and $|\widetilde{RI_{ik}}|$ exchangeable, which will make $S_{ik} = |RI_{ik}| - |\widetilde{RI_{ik}}|$ to follow a distribution symmetric around 0. On the other hand, suppose (x_i, x_j) is a causal relation. Then, S_{ij} will indicate how deviated the relationship of (x_i, x_j) is compared to the null hypothesis, leading to estimation of the effect size. The idea is that knockoff matrix $\tilde{X}$ is generated to mimic the correlation structure within X but minimises the cross-correlation with outcome variable[143]. Specifically, model-X knockoff variables for the set of random variables $X = (x_1, \dots, x_p)^T$ of our interest are a new family of random variables $\tilde{X} = (\widetilde{x_1}, \dots, \widetilde{x_p})^T$ that satisfies two properties: (1) $(X, \tilde{X})_{swap(s)} =^d (X, \tilde{X})$ for any subset $S \subset \{1, \dots, M\}$, where $swap(s)$ means swapping x_j and $\tilde{x_j}$ for each $j \in S$ and $=^d$ denotes equal in distribution, and (2) $\tilde{X} \perp\!\!\!\perp Y|X$, that is, $\tilde{X}$ is independent of X given outcome Y. Suppose $x_j \sim N(0, \Sigma)$ with $\Sigma \in \mathbb{R}^{M \times M}$ the covariance matrix. A valid construction of $\tilde{x_j}$ is

Equation 2.3. $\tilde{x_j}|x_j \sim N\left(x_j - diag\{S\}\Sigma^{-1}x_j, 2diag\{S\} - diag\{S\}\Sigma^{-1}diag\{S\}\right)$.

Model-X knockoffs can be sampled from the conditional distribution of $\tilde{x_i}|x_i$ as follows.

Equation 2.4. $(x_j, \tilde{x_j}) \sim N\left(\binom{0}{0}, \begin{pmatrix} \Sigma & \Sigma - diag\{S\} \\ \Sigma - diag\{S\} & \Sigma \end{pmatrix}\right)$

In this sampling, the sensitivity of identifying x_j can increase with a larger S since it will make the knockoffs more different from S, subjected to another constraint that S should make $\Sigma - diag\{s\} \geq 0$. By pairing knockoff variable $\tilde{x}_j$ with the corresponding input variable x_j and optimizing together for x_i, one can quantify the importance of x_j in reference to $\tilde{x}_j$. Specifically, let $W_i^{(0)} \in \mathbb{R}^{M\times 1}, W_i^{(1)} \in \mathbb{R}^{M\times M}, W_i^{(2)} \in \mathbb{R}^{M\times M}, and\ W_i^{(3)} \in \mathbb{R}^{M\times 1}$ be the weight matrices connecting the input vector to the first hidden layer, the first hidden layer to the second hidden layer, the second hidden layer to the third hidden layer, and the third hidden layer to x_i, respectively. The weight estimates can be summarized into $w_i = W_i^{(0)} \otimes (W_i^{(1)} W_i^{(2)} W_i^{(3)})$, where $\otimes$ denotes the element-wise matrix operation. Also, let ri_{ji} and $\widetilde{ri_{j\iota}}$ be the filter weight for x_j and its knockoff counterpart $\tilde{x}_j$. Then, variable importance values can be estimated for input and knockoff variables as follows.

Equation 2.5. $RI_{ji} = ri_{ji} \times w_i\ and\ \widetilde{RI_{j\iota}} = \widetilde{ri_{j\iota}} \times \widetilde{w_\iota}$.

We use Adam to train this deep learning model with respect to the mean squared error loss, using an initial learning rate of 0.001 and batch size 10. With $S_{ji} = \left|RI_{ji}\right| - |\widetilde{RI_{j\iota}}|$, DAG-deepVASE estimates effect size on $(x_j \in X_{\setminus i},\ x_i)$ adopted from [106], [144], which can be described in the following two options:

Equation 2.6. $T = min\left\{t \in S, \frac{|\{j:S_{ji} \leq -t\}|}{|\{j:S_{ji} \geq t\}|} \leq q\right\} or\ T_+ = min\left\{t \in S, \frac{1+|\{j:S_{ji} \leq -t\}|}{1 \vee |\{j:S_{ji} \geq t\}|} \leq q\right\}$

where q is a user-defined nominal false discovery rate and T or T_+ is a threshold value for determining which features should be selected. We controlled FDR $q = 0.05$ based on S_{ji}. While this setting has previously been used for a variable selection problem with respect to outcome [31], we extend this problem to estimate the nonlinear effect size for associated variables in this project. Since model-X knockoff assumes to follow Gaussian distribution, we will include only continuous

or ordinal categorical variables that approximately follow Gaussian distribution (using Q-Q plot). We set parameters of DAG-deepVASE according to a guideline that utilized the knockoff framework for variable selection [131] (**Table 3.1**).

In the third step, for each identified variable association (x_i, x_j), whether linear or nonlinear, DAG-deepVASE determines the causal direction as extended from the DG framework as follows. calculated as:

Equation 2.7. $DG(G,Z) = \sum_{j=1}^{p} dg\left(Z_j | Z_{Pa_j^G}\right)$,

where

$$dg\left(Z_j | Z_{Pa_j^G}\right) = \ell\left(\hat{\theta}_{mle} | Z_{\{j\}\cup Pa_j^G}\right) - \ell\left(\hat{\theta}_{mle} \middle| Z_{Pa_j^G}\right) - \frac{c}{2}|Z_j| \left|Z_{Pa_j^G}\right| \log(n),$$

where c is a penalty discount used to tune the density of the resulting graph. Also, $\ell(\hat{\theta}_{mle}|Z_{sub})$, which is the log-likelihood of a subset of Z, is computed using the Gaussian log-likelihood function in reference to $\widehat{\sum}_{sub}$, the partial covariance matrix for the input variables. Note $dg\left(Z_j | Z_{Pa_j^G}\right) = logP\left(X_j | X_{Pa_j^G}\right)$ if the data has only continuous variables[36]. By maximum likelihood, the DG framework determines x_j as causal and x_i as effect if $dg(x_i|x_j) > dg(x_j|x_i)$ or $dg(x_i|x_j) - dg(x_j|x_i) > 0$. Due to multiplication commutativity,

Equation 2.8. $dg(x_i|x_j) - dg(x_j|x_i)$

$$= \ell(\hat{\theta}_{mle}|x_{\{i,j\}}) - \ell(\hat{\theta}_{mle}|x_j) - \frac{c}{2}|x_i||x_j|\log(N) - \left(\ell(\hat{\theta}_{mle}|x_{\{j,i\}}) - \ell(\hat{\theta}_{mle}|x_i) - \frac{c}{2}|x_j||x_i|\log(\mathrm{N})\right)$$

$$= \ell(\hat{\theta}_{mle}|x_i) - \ell(\hat{\theta}_{mle}|x_j)$$

$$= l\left(\frac{\hat{\theta}_{mle}|x_i}{\hat{\theta}_{mle}|x_j}\right). (1)$$

After running this likelihood ratio test, we algorithmically remove the causal relations that create a cycle (a non-empty tail in which the first and the last nodes are equal) to ensure acyclicity by removing the one association with the least effect size (S_{ji}).

3.3.10 Simulation for Nonlinear Associations

The nonlinear simulation datasets were generated using a single index model [145]–[148]. Each simulation dataset consists of three parts: outcome variable $y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^{N\times 1}$; a set of independently and identically distributed random variables $X \in \mathbb{R}^{N\times M}$ which have a different degree of nonlinear association with y; and a set of independently and identically distributed random variables $Z \in \mathbb{R}^{Q\times M}$ which have no association with Y. The following model was used to generate associated variable pairs x_i and Y:

Equation 2.9. $Y_i = \alpha g(x_i^T \beta) + (1-\alpha) x_i^T \gamma + \varepsilon_i$

where g is a nonlinear link function which we set to be a cube (X^3) function, Y_i is the outcome value and ε_i is noise added to the ith outcome. α determines the proportion of the (non)linearity of the simulation where $\alpha = 1$ determines the association of X_i and Y_i only with the nonlinear link function (complete-nonlinear) and $\alpha = 0.5$ determines the association half by the nonlinear function and half by the linear function (partial-nonlinear). The distribution for noise ε was simulated from $\mathcal{N}(0, \sigma^2 \mathrm{I}_N)$, where σ is set as 1. The rows of X was simulated independently from a distribution $\mathcal{N}(0, \sum)$ with a precision matrix $\sum^{-1} = (\rho^{|j-k|})_{1\le j,\ k\le (q+p)}$ with $\rho = 0.5$. A similar strategy has been used to assess the performance of deep-learning methods developed for variable selection and causal inference, deepPINK[131], and DAG-GNN[37], respectively. In this project, we extended their methods by diversifying the degree of nonlinearity by adding the

proportion of linearity α. Also, note that this simulation satisfies the essential condition for causal inference, causal sufficiency. Specifically, Y is the direct product of X without a mediator (**Equation 2.9**). Since this means no latent confounder in the causal relationship from X to Y, it satisfies causal sufficiency. The other two essential causal assumptions, causal Markov and faithfulness are not relevant to the simulation since there is no other variable in the simulation that is conditionally dependent or independent of X and Y. Altogether, this simulation experiment is designed to evaluate the performance of causal inference methods in a straightforward setting.

For each parameter combination (number of features and samples, complete- or partial-nonlinear), we ran various numbers of repetitions (50, 100, and 150), but report the results of 50 repetitions as different numbers of repetitions returned very similar results.

3.4 Results

3.4.1 DAG-deepVASE Improves Power in Identifying Nonlinear Causal Relations in Simulation Data

To evaluate the performance of DAG-deepVASE in the presence of multiple causal variables, we compared DAG-deepVASE with competing methods on simulation data. Such methods include causalMGM, DG, NOTEARS, and DAG-GNN. We included causalMGM and DG because they employ the two-step strategy as DAG-deepVASE: identifying variable associations and then learning the causal direction of the associations. While causalMGM was

originally developed with the two-step strategy, DG does not have the first step because DG is developed to learn causality based on given associations. To be fair to DG, we developed the first step for DG: in the first step, we applied MGM to identify associations and in the second step, we used the original DG to learn their causalities. We will refer to this model as linear-DG model since the MGM implementation identifies variable associations based on the linear interaction terms. Also, note that whereas DAG-deepVASE identifies both linear and nonlinear associations and uses DG to learn their causal directions, linear-DG identifies linear associations and uses DG to learn their causal directions, and causalMGM identifies only linear associations and applies PC to learn their causal directions. We included NOTEARS and DAG-GNN because they are established DNN methods to infer causality. We ran the methods using default parameters or those suggested by the authors throughout this project (**Table 3.1**).

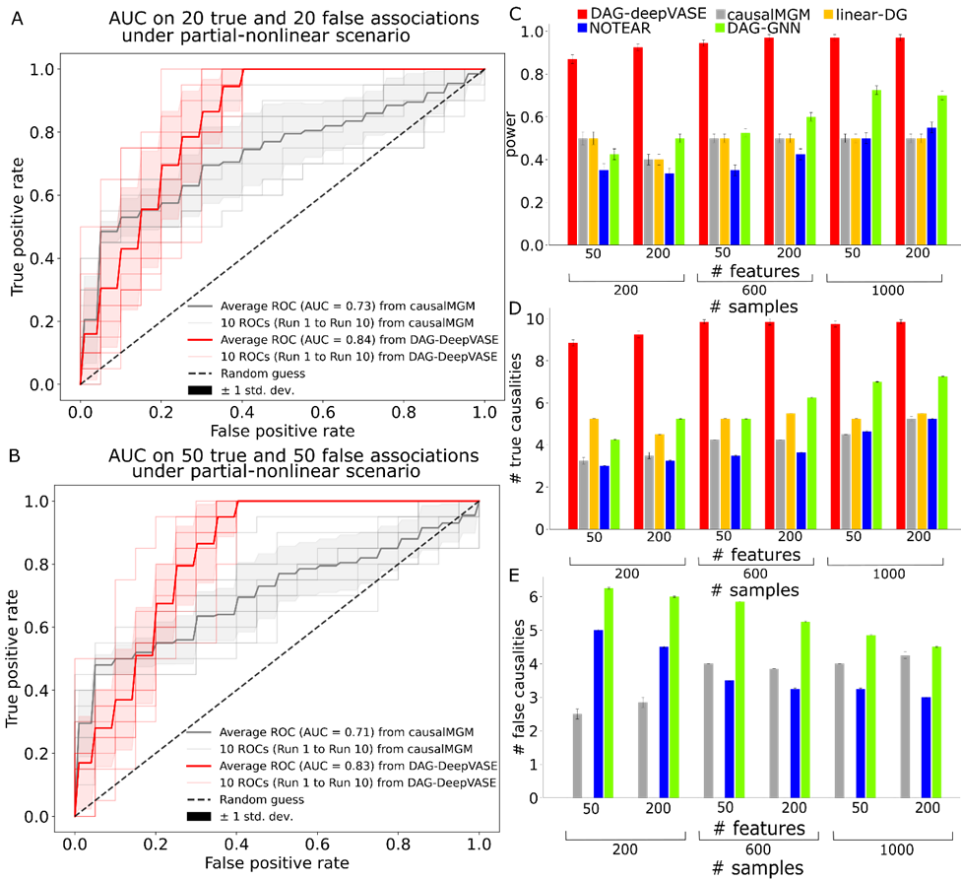
To compare the methods in sensitivity and specificity simultaneously, we simulated 10 data sets of 40 or 100 variables where half (20 or 50, respectively) of the variables collectively determine the outcome (true associations) and the other half are not associated with the outcome (false associations, see **Materials and Methods**). Each data set was simulated for 10,000 samples. To mimic biological variables that would interact in various degrees of nonlinearity, simulations were conducted under two scenarios: complete-nonlinear or partial-nonlinear scenarios. We ran DAG-deepVASE and causalMGM on the datasets. We did not run linear-DG since it identifies the same association pairs as causalMGM. We did not run NOTEARS and DAG-GNN for this experiment since it is not straightforward to vary threshold values for plotting the ROC curve in the DNN architecture. In both complete- and partial-nonlinear scenarios, DAG-deepVASE consistently outperformed causalMGM in AUC (area under the receiver operating characteristic curve). Specifically, for the simulations with 40 and 100 associations under the complete-nonlinear

scenario, DAG-deepVASE achieves an average of 0.84 and 0.82 AUC, respectively, outperforming causalMGM which achieves an average of 0.71 and 0.68 AUC (**Figure. 3.2A** and **Figure. 3.2B**, respectively). The same trend is observed under the partial-nonlinear scenario where DAG-deepVASE achieves an average of 0.84 and 0.83 AUC and causalMGM achieves an average of 0.73 and 0.71 AUC for the simulations with 40 and 100 associations (Supplemental Figure. 6**A** and Supplemental Figure. 6**B**, respectively).

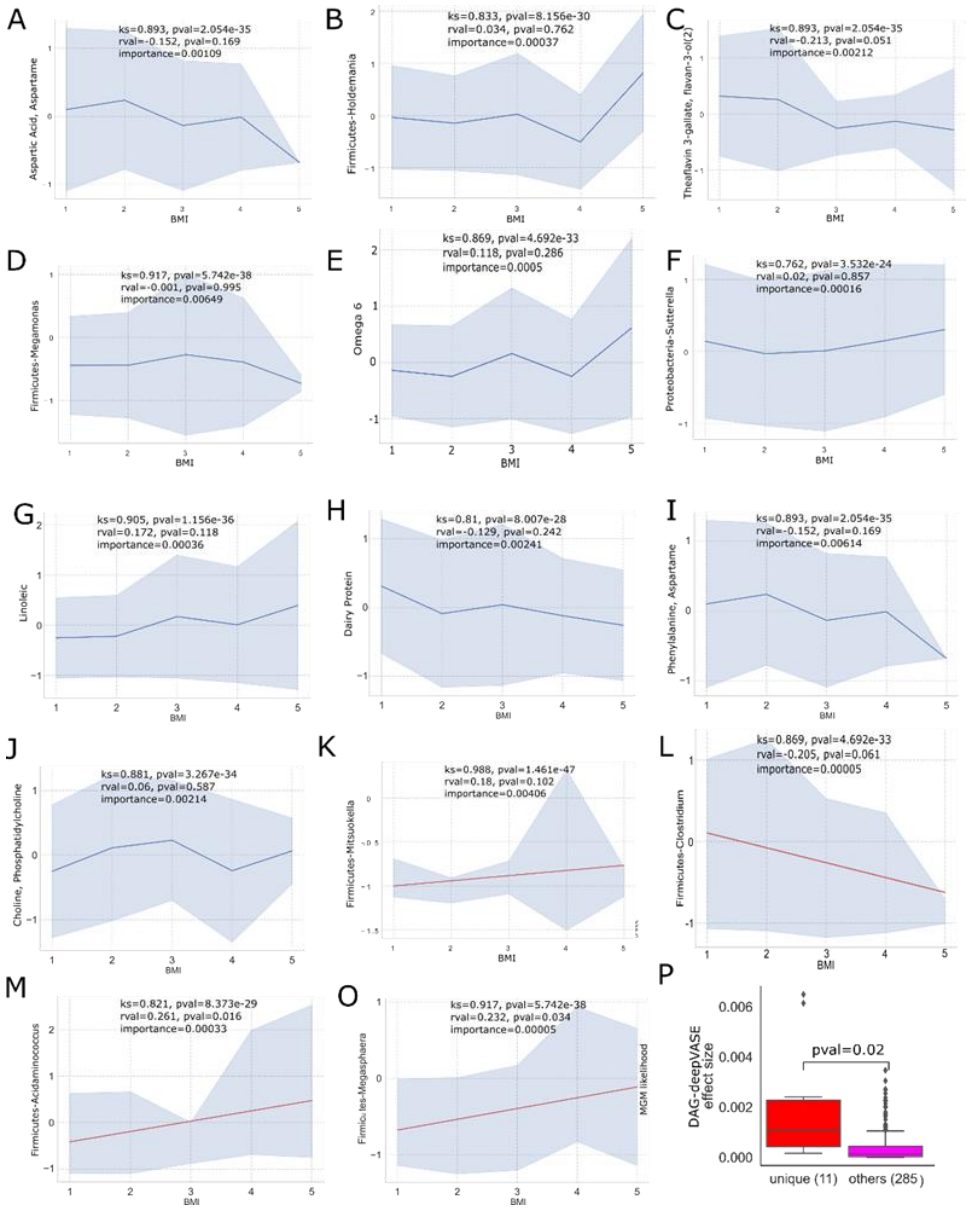
To further mimic biological situations where true associations would be relatively rare among all pairwise combinations of biological variables, we simulated different numbers of variables (M = 50, 100, 200, 400, 600, 800, 1000, 1500, 2000, 2500, and 3000) with various sample sizes (N=200, 600, 1000), where ten variables collectively determine the outcome (true associations). For each combination of variable number and sample size, we conducted the simulation experiment 50 times. In the complete-nonlinear simulation scenario, we first compared the number of true associations identified by each method before assessing the causal directions. DAG-deepVASE shows a two-fold higher power than the other methods by identifying more than 90% of the true associations in most simulation scenarios (**Figure. 3.2C**). Interestingly, while DAG-GNN performs slightly better than linear approaches, causalMGM and linear-DG, in terms of power and sensitivity, NOTEARS performs the worst in all scenarios in general. Second, we compared the number of true and false causal directions learned from the identified associations (**Figure. 3.2D,** respectively**, Supplemental Table. 2. Tab 1**). In all experiments under the complete-nonlinear scenario, DAG-deepVASE consistently outperforms the other methods in identifying true causalities. Especially, for larger sample sizes (n=600 and 1,000), DAG-deepVASE identified more than 97% of the true causalities. causalMGM returned bidirectional causal directions for all identified associations, which are counted as both true and false positives.

On the other hand, although linear-DG identified less than half of the true associations as mentioned above, it learned the true causalities on the small number of the identified associations (**Figure. 3.2D**), demonstrating that DG can be used to learn nonlinear causalities. Further, the other DNN methods also identified less than 50% of the true causalities than DAG-deepVASE. Together with such high true positive rates, DAG-deepVASE also outperforms the other methods by not identifying any false casualties in any of the scenarios, whereas competing methods suffer from high false causalities. For example, causalMGM returns 3~5 false-positive causalities by returning bidirectional causalities (**Figure. 3.2E, Supplemental Figure. 5D**) and both DNN methods, NOTEARS and DAG-GNN, suffer from the highest number of false causalities. In the partial-nonlinear scenario, a very similar result is returned for power (**Supplemental Figure. 5C)**, true

positive causalities (**Supplemental Figure. 5D)**, and false positive causalities (**Supplemental Figure. 5E**).



Supplemental Figure 5. AUC estimated for DAG-deepVASE and causalMGM on (A) 20 true and false associations and (B) 40 true and false associations, both under complete-nonlinear scenarios. (C) Average number, and standard error (error bar), of true associations in the partial-nonlinear scenario identified by DAG-deepVASE (red), causalMGM (gray), linear-DG (yellow), NOTEAR (blue), and DAG-GNN (green) over 50 runs in various simulation scenarios, varying the number of features and sample sizes. Average number, and standard error (error bar), of (D) true causalities and (E) false causalities over 50 runs. DAG-deepVASE and linear-DG did not identify any false causalities.



Supplemental Figure 6. (A-J) Variable values against the BMI value window (1~5) that are identified as a nonlinear association to BMI (K-O) Variable values against the BMI value window (1~5) that are identified as a linear association to BMI. In the figures, KS is Kolmogorov-Smirnov (KS) test statistic, p-value is estimated from the KS test, rval is from a linear regression model, pval is from the linear regression, and importance is measured in DAG-deepVASE. The gray area indicates 95% confidence intervals, the blue line indicates median values, and the red line represents a linearly regressed line. P-value for linear fit is calculated from a permutation test with R2 (Methods) (P) Effect size estimated by DAG-deepVASE for 16 validated factors and 285 other factors to BMI.

Altogether, DAG-deepVASE outperforms the other methods by identifying the highest number of true nonlinear associations and by learning the highest number of true causalities across various simulation scenarios without false positives, while competing methods could identify less than half of true nonlinear causalities with several false positives.

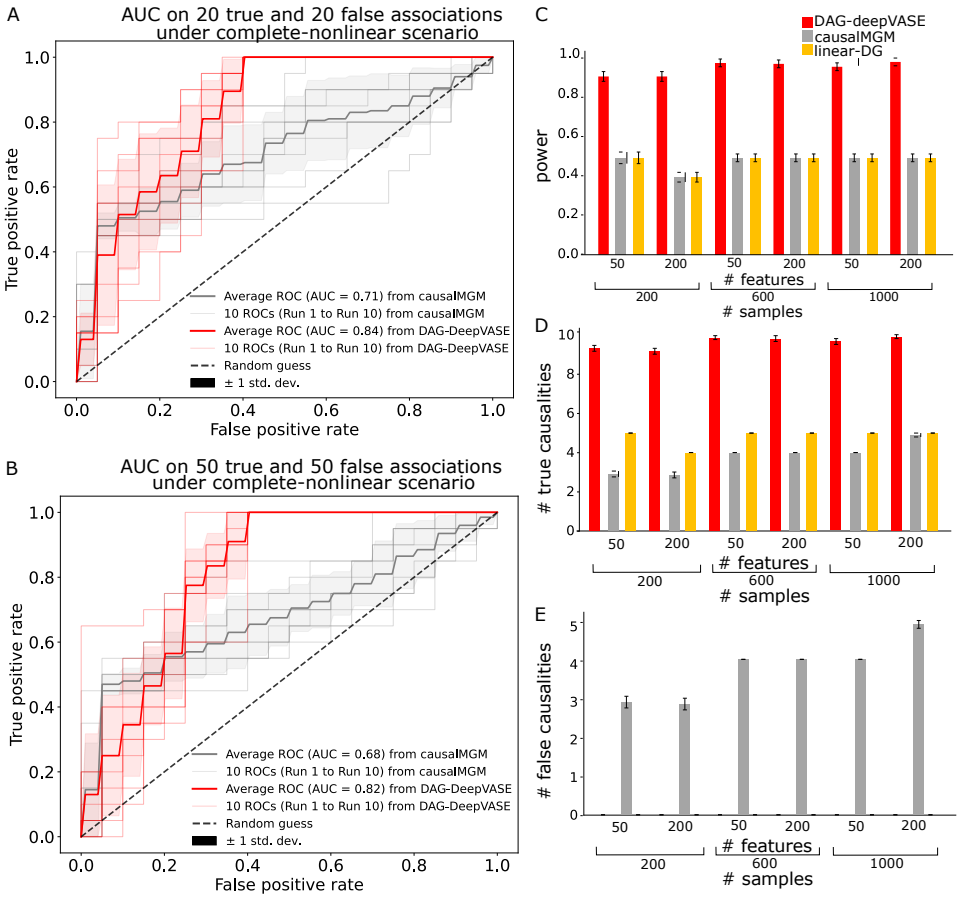


Figure 3.2. Performance assessment of causal inference methods on the simulated data AUC estimated for DAG-deepVASE and causalMGM on (A) 20 true and false associations and (B) 40 true and false associations, both under complete-nonlinear scenarios. (C) Average number, and standard error (error bar), of true associations in the complete-nonlinear scenario identified by DAG-deepVASE (red), causalMGM (gray), linear-DG (yellow), NOTEAR (blue), and DAG-GNN (green) over 50 runs in various simulation scenarios, varying the number of features and sample sizes. Average number, and standard error (error bar), of (D) true causalities and (E) false causalities. DAG-deepVASE and linear-DG did not identify any false causalities.

3.4.2 DAG-deepVASE Identifies Both Linear and Nonlinear Associations Among Clinical Features With High Sensitivity in Pediatric Sepsis Data

To demonstrate the importance of identifying nonlinear variable associations for sensitive causal inference, we first focus on identifying associations among diverse types of variables in clinical data. The data consists of clinical and biomarker variables (laboratory parameters, cytokines, and chemokine measurements) from 404 children with severe sepsis[149]. We compared DAG-deepVASE and causalMGM in this section. We excluded linear-DG because they identify the same set of associations with causalMGM. We excluded NOTEAR and DAG-GNN from further analyses since they identified high rates of false positive causalities in simulation studies. Since DAG-deepVASE assumes that the variables follow the Gaussian distribution, we consider 45 continuous or ordinal categorical variables excluding one binary/nominal variable in the data set. Among the variables, DAG-deepVASE identifies 118 associations (**Figure 3.3A**), whereas causalMGM identifies 42 associations (49.5%, **Supplemental Table. 2. Tab 2**) of the associations. Since causalMGM is only able to identify linear associations, the 42 associations are likely linear. Many of the identified linear associations are already clinically and biologically verified. For example, the serum level of soluble CD163 (sCD163), a macrophage activator[150], only has linear associations (**Figure 3.3B**) with biomarkers known to activate macrophages, such as M-CSF (macrophage colony-stimulating factor)[151], MCP-1 (monocyte chemoattractant protein-1)[152], Il-1b[153], TNF-a [154], [155] and other key drivers of macrophage response including its ligand hemoglobin [156]. Also, age is another variable only linearly associated with other variables, including heart rate, creatinine, and lymphocyte count (**Figure 3.3B**). Since each

of them changes monotonically with age in pediatric subjects [157]–[159], it is reasonable that they are identified as linear associations.

In addition to the 42 linear associations that are identified by both DAG-deepVASE and causalMGM, DAG-deepVASE uniquely identifies 76 nonlinear associations. Multiple nonlinear associations have been validated in previous clinical and biological studies with an implication for nonlinearity. An example is an association between systemic inflammatory response syndrome (SIRS) status and heart rate (**Figure 3.3B**). This association is expected as nonlinear, as the SIRS status is diagnosed by a nonlinear combination, which is the presence of any two of the four clinical criteria, including tachycardia (elevated heart rate) [160]. Also, as the SIRS response is defined as a result of systemic immunological activation, DAG-deepVASE uniquely found nonlinear associations between SIRS status and pro-inflammatory cytokines including CRP [161], [162], IL-1β [163], and IFN-γ [164] (**Figure 3.3B**), corroborating their collective roles in inflammation. Since cytokines are produced involving different combinations of signal transduction pathways [165], [166], their associations with SIRS are expected to be nonlinear rather than linear.

While the method identified validated associations, DAG-deepVASE also identified novel nonlinear relationships of clinical potential for future validation. For example, it identified the nonlinear associations between central nervous system (CNS) dysfunction and SIRS and between IL-22 and SIRS (**Figure 3.3B**). The former is validated: critically ill patients with SIRS are known to have a measurable risk for organ dysfunction such as CNS dysfunction [167], [168]. This validation also confirms our causal inference that found the causal direction from SIRS status to CNS dysfunction (**Supplemental Table. 2. Tab 2**). As it is imperative to elucidate how SIRS interacts with modifiable cytokines for clinical potential, our causal inference also suggests the novel clinical potential of IL-22 to treat CNS dysfunction through modifying SIRS status. While

IL-22 plays a key role in immunoregulation and has been linked to the development of organ failure in mouse models of abdominal sepsis [169], DAG-deepVASE revealed the causal relationship from IL-22 to SIRS status in children with sepsis. Especially, it identified this relationship by strong effect size (top 19th out of 118, **Supplemental Table. 2. Tab 2**), suggesting a strong reproducibility and thus clinical utility. After more experimental validations, this result can help design future clinical trials to treat organ dysfunctions with IL-22 for pediatric sepsis. Altogether, DAG-deepVASE identifies both validated and novel findings by identifying linear and nonlinear associations with high sensitivity.

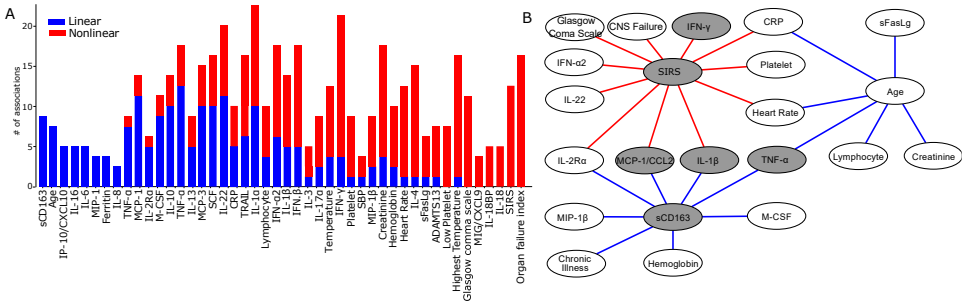


Figure 3.3. Linear and nonlinear associations in pediatric sepsis data (A) Number of linear (blue) and nonlinear (red) associations involving each of the 45 variables. (B) A subnetwork of linear (blue) or nonlinear (red) variable associations involving SIRS (associated only non-linearly) and sCD183 (associated only linearly) and with normalized effect size. Gray nodes connect between IFN-γ and TNFα. For full names of the variables, readers are referred to Materials and Methods.

3.4.3 DAG-deepVASE Accurately Identifies Nonlinear Causalities and Estimates Their Effect Sizes in the Nutrients/Gut Bacteria and Body-Mass Index (BMI) Data.

Variables in complex biological systems interact in varying degrees of nonlinearity [170]–[172]. To examine the sensitivity of DAG-deepVASE in the presence of various degrees of nonlinearity, we compared DAG-deepVASE with causalMGM and linear-DG on a cross-sectional

data set consisting of 214 nutrient intakes, 87 gastrointestinal (GI) bacteria genera and body-mass index (BMI) collected from 90 healthy volunteers[132]. Note that nutrient intakes would affect GI bacteria before affecting BMI, suggesting generally a more nonlinear relationship between the nutrient intakes and BMI than between them and GI bacteria. For a balanced assessment, we selected the same number (8) of nutrient intakes and bacteria genera that are known to affect BMI in animal experiments or clinical trials out of the 214 nutrient intakes and 87 bacteria genera data (**Supplemental Table. 2. Tab 3**). We selected the 16 features also because they were previously suggested to have nonlinear associations with BMI by a DNN-based variable selection method [131]. DAG-deepVASE identified 15 associations, while causalMGM and linear-DG identified only 5 associations (31.3%): all these 5 associations are between specific GI bacteria and BMI (**Figure 3.4A**). Note that causalMGM and linear-DG fail to identify any association between nutrient intakes and BMI, while DAG-deepVASE could identify all 8 of them. Since nutrient intakes likely affect BMI more nonlinearly than between GI bacteria and BMI, this result reaffirms that DAG-deepVASE uniquely identifies nonlinear relationships. To characterize the nonlinear associations, we examined how the 8 nutrient intake and 8 bacteria genera levels change against the BMI value. The 5 associations between GI bacteria and BMI identified by all three methods show a single linear association throughout the BMI region (**Figure 3.4B** (p-value for linear fit: 0.001)**, Supplemental Figure. 6K-O**). On the other hand, the other 8 associations between nutrient intakes and BMI and 3 associations between GI bacteria and BMI, which are identified by only DAG-deepVASE, show nonlinear relationships (**Figure 3.4C** (p-value for linear fit: 0.58), **Supplemental Figure. 6A-J** (p-value for linear fit on average 0.42)), characterized by multiple sub-trends across the BMI ranges. For example, choline and phosphatidylcholine w/o suppl. intake

(**Figure 3.4C**) shows an increasing trend from BMI 1~3, a decreasing trend from BMI 3~4, then another increasing trend from BMI 4~5.

In the second step of determining causalities from the identified associations, we deemed true the causal directions from nutrient intake/ bacteria genera to BMI based on literature in **Table 3.3**. DAG-deepVASE identified true causal directions from all the 15 associations it found. On the other hand, as causalMGM uses PC to learn the causal directions of the associations, PC removed 2 of the 5 associations in its step of testing the conditional independence relationship and identified the other 3 associations as bidirectional causalities that we considered to be both false positive and false negative (**Figure 3.4D**). While linear-DG identified 5 true causal directions on the five identified associations, it still could not learn 11 causalities because of its inability to identify nonlinear causality. Altogether, DAG-deepVASE outperforms other methods due to its ability to identify nonlinear associations combined with the excellent performance of DG in learning nonlinear causalities among the identified associations.

Table 3-3. nonlinear associations (8 nutrient intakes and 8 bacteria genera) that were validated in literature.

Nutrient intake		Bacteria genera		
Micronutrient	Reference	Phylum	Genus	Reference
Linoleic Acid, Aspartame	[173]	Proteobacteria	Sutterella	[174]
Omega 6	[175]	Firmicutes	Allisonella	[135]
Dairy Protein	[176]	Firmicutes	Holdemania	[177]
Aspartic	[178]	Firmicutes	Mitsuokella	[179]

Phenylalanine, Aspartame	[178]	Firmicutes	Clostridium	[135]
Choline, Phosphatidylcholine	[180]	Firmicutes	Megamonas	[181]
Theaflavin 3-gallate, flavan-3-ol(2)	[182]	Firmicutes	Megasphaera	[179]
Choline, Phosphatidylcholine w/o suppl.	[180]	Firmicutes	Acidaminococcus	[135]

Further, to demonstrate how the effect size DAG-deepVASE estimates leads to the unbiased discovery of causal relations, we estimated the effect size of all 301 potential associations between nutrient intake/bacteria genera and BMI using DAG-deepVASE including non-validated ones (**Supplemental Table. 2. Tab 3**). The top 5 associations that have the largest effect sizes estimated by DAG-deepVASE's nonlinear module are all validated: Meganomas [181], Phenylalanine [178], Mitsuokella [179], Parvimonas [178] and Sporobacter [179]. Since these findings were independent, it is difficult to prioritize their importance. To conduct further experimental validations or clinical trials that target a limited number of strong associations, estimating effect sizes via DAG-deepVASE enables to prioritize important variables. Also, while

the 5 top features are already validated, it would be interesting to validate other novel features with large effect sizes.

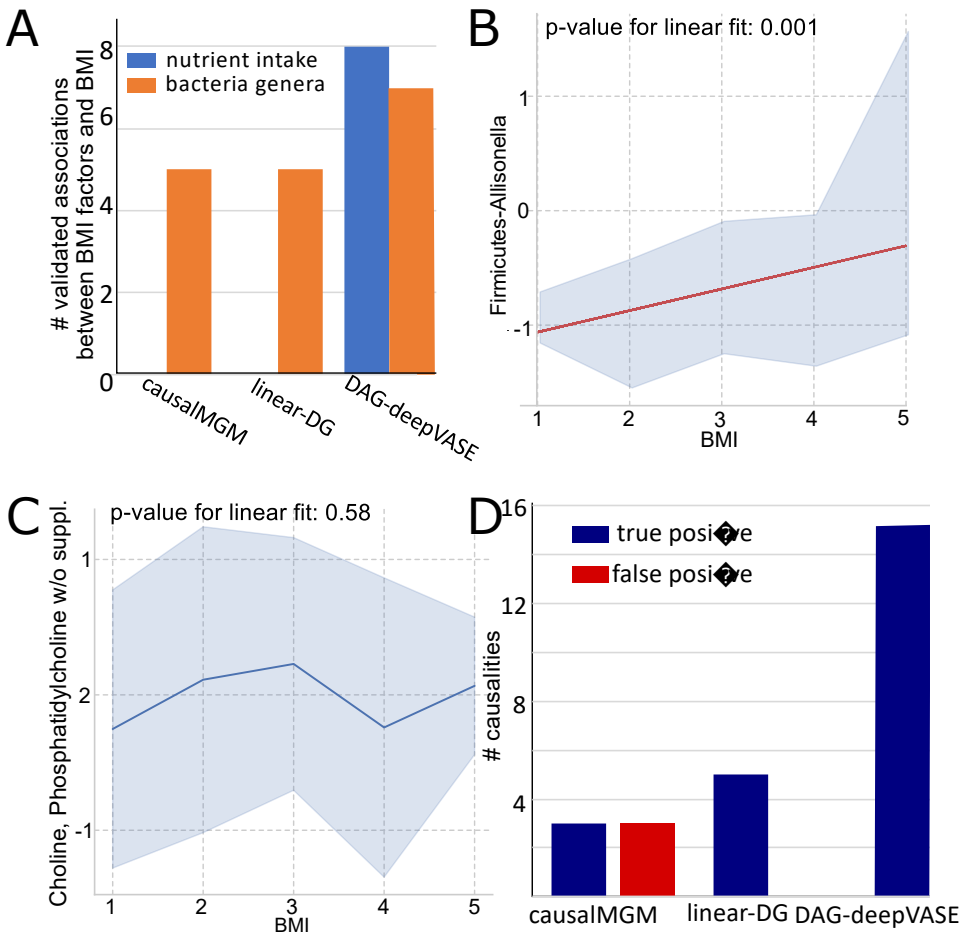


Figure 3.4. Performance assessment of four causal inference methods on various degrees of nonlinear associations in BMI/bacteria/gut microbiome data (A) Number of associations the methods (causalMGM, linear-DG, and DAG-deepVASE) identified between the BMI status and 8 nutrient intake (blue) and 8 bacteria genera in the gut (red) that are validated associated with the BMI status. (B) The relationship between BMI and Firmicutes-Allisonella identified with confidence interval (gray intervals). Red line represents the estimated linear regression and p-value for linear fit is calculated from a permutation test with R^2 (Materials and Methods). (C) The relationship between BMI and Choline, Phosphatidylcholine w/o suppl identified with confidence interval (gray intervals). Blue line connects the middle point of the BMI values 1 to 5. (D) Number of true positive (dark blue) and false positive (red) causalities identified by causalMGM, linear-DG, and DAG-deepVASE. DAG-deepVASE and linear-DG did not identify any false causalities.

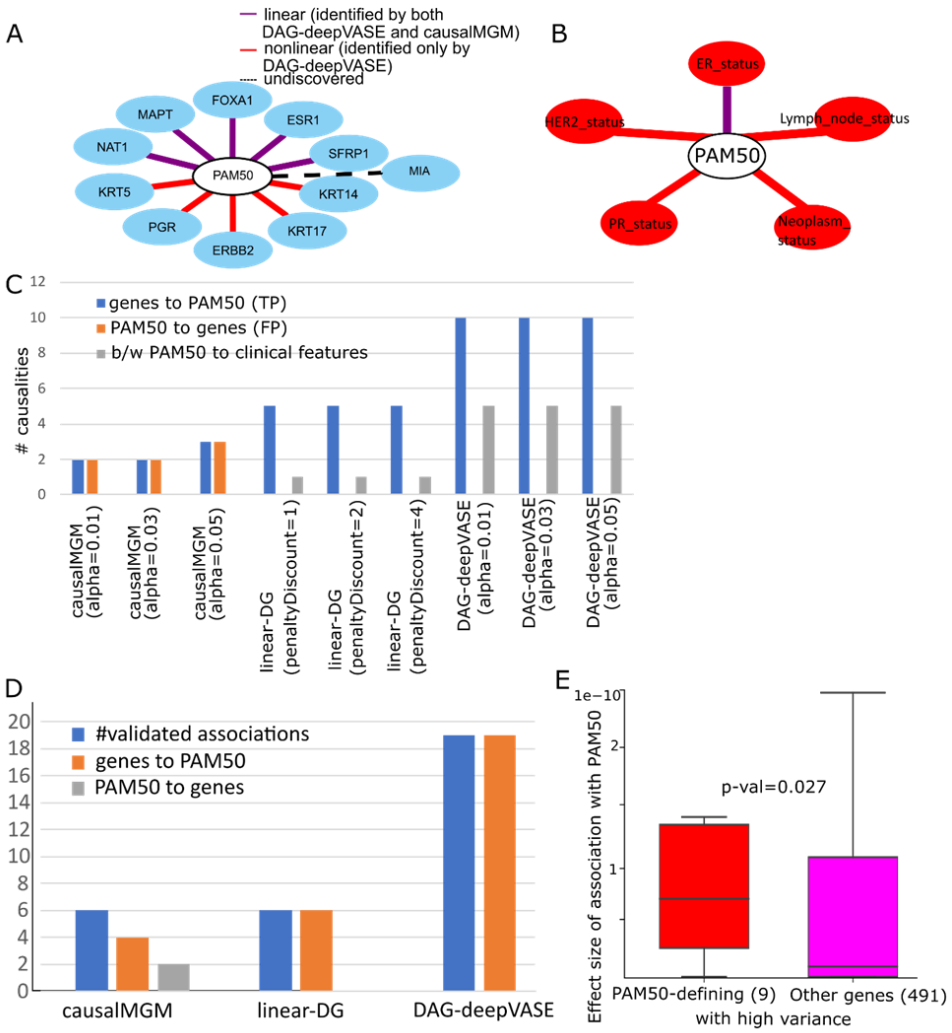
3.4.4 DAG-deepVASE Identifies Causal Relations Among Molecular and Clinical Variables in Breast Cancer Data.

Among various types of variable interactions in a complex disease, identifying causal relationships between molecular variables (e.g., gene expression) and clinical variables (e.g., cytokine measurements in the serum) are particularly interesting because the findings can help identify molecular therapeutics. To evaluate the performance of DAG-deepVASE in learning the complex molecular pathogenic mechanisms, we compared DAG-deepVASE with causalMGM and linear-DG on the TCGA breast cancer of gene expression and clinical variables, such as PAM50 (n=601 tumor samples). PAM50 is an important clinical feature to categorize breast tumors, which is defined by the tumor's expression of 50 genes (PAM50 genes) [183]. Therefore, we consider the causal directions from the genes to the PAM50 status as true positives. For our analysis, we chose 10 of the 50 genes (PAM50-defining genes) that are also included in the top 500 genes that have the highest variance across the samples (high variance gene set). In addition to the 10 genes, we also considered 5 clinical variables, which are known to characterize breast tumors with the PAM50 status: estrogen receptor (ER) [184], progesterone receptor (PER) [184], human epidermal growth factor receptor (HER)[185], lymph node status[186], and tumor staging code[187].

In the first step of identifying associations between the 10 PAM50-defining genes with the highest variances and PAM50 status, DAG-deepVASE identified 9 associations out of 10 associations, while causalMGM and linear-DG identified only 5 of them, attributing the 40% power increase of DAG-deepVASE to the identification of nonlinear associations. Between the 5 clinical variables and PAM50, DAG-deepVASE identified all 5 associations while causalMGM and linear-DG identified only one association (20%) (**Figure 3.5A, Supplemental Figure. 7A,**

B), suggesting that 80% (4 of 5) of the associations are nonlinear. In the second step of learning causal directions, DAG-deepVASE outperforms both causalMGM and linear-DG, identifying true causalities from all 9 identified associations between the genes and PAM50 (**Figure. 3.5B**). On the other hand, causalMGM identified bidirectional causalities for 3 associations after the PC step removed the other two associations based on the conditional independence relationships. And linear-DG identified the correct causalities on all 5 identified associations but still missing causalities for the other 5 associations that linear-DG could not identify. We did not assess the causal directions between the 5 clinical variables and PAM50 since the true causal directions are not clear between them. We tried different parameter settings of the methods to find that this trend

holds true across the parameter settings (**Supplemental Figure. 7C, Supplemental Table. 2. Tab 4**).



Supplemental Figure 7. (A) PAM50-defining genes associated with the PAM50 status of patients identified by both DAG-deepVASE and causalMGM (purple) or uniquely by DAG-deepVASE (red). Both DAG-deepVASE and causalMGM could not identify a PAM50-defining gene (MIA) in a dotted line. (B) clinical features (e.g., hormone status) of the breast cancer samples associated with PAM50 that are identified by both DAG-deepVASE and causalMGM (purple) or uniquely by DAG-deepVASE (red). (C) Number of causalities identified by causalMGM, linear-DG, and DAG-deepVASE run with various parameter settings. (D) Number of validated associations (blue), causalities identified from genes to PAM50 status (orange) or from PAM50 status to genes (gray) when 20 PAM50 genes are run on DAG-deepVASE. (E) Effect size

estimated by DAG-deepVASE for 9 PAM50-defining genes and 491 other genes in the 500 genes with highest expression variance.

To ensure reproducibility of this finding, we further selected the top 20 genes from PAM50 genes with the largest variance in the data and evaluated the methods on the genes. In identifying the true associations, DAG-deepVASE identified 19 associations out of 20 (95.5%) while both causalMGM and linear-DG identified 6 of them (30%) (**Supplemental Figure. 7D**). And, in learning the causal directions, DAG-deepVASE identified 19 true causalities from all the identified associations, while causalMGM identified 4 true and 2 false causalities out of the 6 associations and direct-DG identified 6 true causalities from all the identified associations (**Supplemental Figure. 7D**). Altogether, the results demonstrate that DAG-deepVASE outperforms causalMGM and linear-DG in identifying true associations, learning true causalities, and differentiating false causalities in the breast cancer data.To demonstrate how DAG-deepVASE enables us to understand complex pathogenic mechanisms across multiple regulatory layers in breast cancer, we expanded our analysis by investigating causalities among the 10 genes and the 6 clinical features, including the PAM50 status. Specifically, we inspected whether it is linear or nonlinear causalities in the following categories: causalities between a gene and a clinical feature and causalities between clinical features. First, while only a few causalities between genes were identified as non-linear interactions (2 of 8 (25%)), most of the causalities between clinical features and between a clinical feature and a gene were identified as nonlinear causalities (13 out of 15 (86.7%) and 34 out of 43 (79.1%) respectively). While many of them are previously validated in clinical trials or biological experiments, e.g., from ERBB2 (HER2) to PR (progesterone receptor), ERBB2 to ER (estrogen receptor), and ERBB2 to PAM50 [188]; KRT5 (keratin5) to PR and KRT5 to ER[189] (**Figure. 3.5D**), the prevalence of nonlinear causality is consistent with the expectation since the

clinical features, mostly hormone receptor status, are regulated through multiple biochemical pathways[26], and thus these relations are likely nonlinear. Other studies also advocate the nonlinear interactions of the clinical features by showing that incorporating nonlinearity in statistical models improves the prediction accuracy of their effects on breast tumor biology, e.g., in the transcriptional profile and survival analysis[23]–[25]. Second, between a gene and a clinical feature, the method found that all 43 causal directions are from genes to clinical features (**Figure. 3.5C**). Since the clinical features in this data are mainly hormone receptor status, this result conforms to the expectation that genes code for the hormone receptor activity[190]. Incorporating all linear and nonlinear causalities under the categories sheds mechanistic insight into the complex tumorigenic process underlying breast cancer. For example, although the keratin genes (KRT5, KRT14, and KRT17) were found to interact in cancer genome studies [191]–[193], it was not clear how the cluster affects clinical features for cancer. Our result suggests that it is because the gene cluster is formed in linear interactions but its effect on clinical features is mostly nonlinear. Altogether, DAG-deepVASE could identify nonlinear causalities, consisting of 74.2% of all causalities in this data, which would be missed by other existing methods. Identifying nonlinear

causal relationships sheds insights into not only genetic interactions but also their interactions with clinical features of tumor biology.

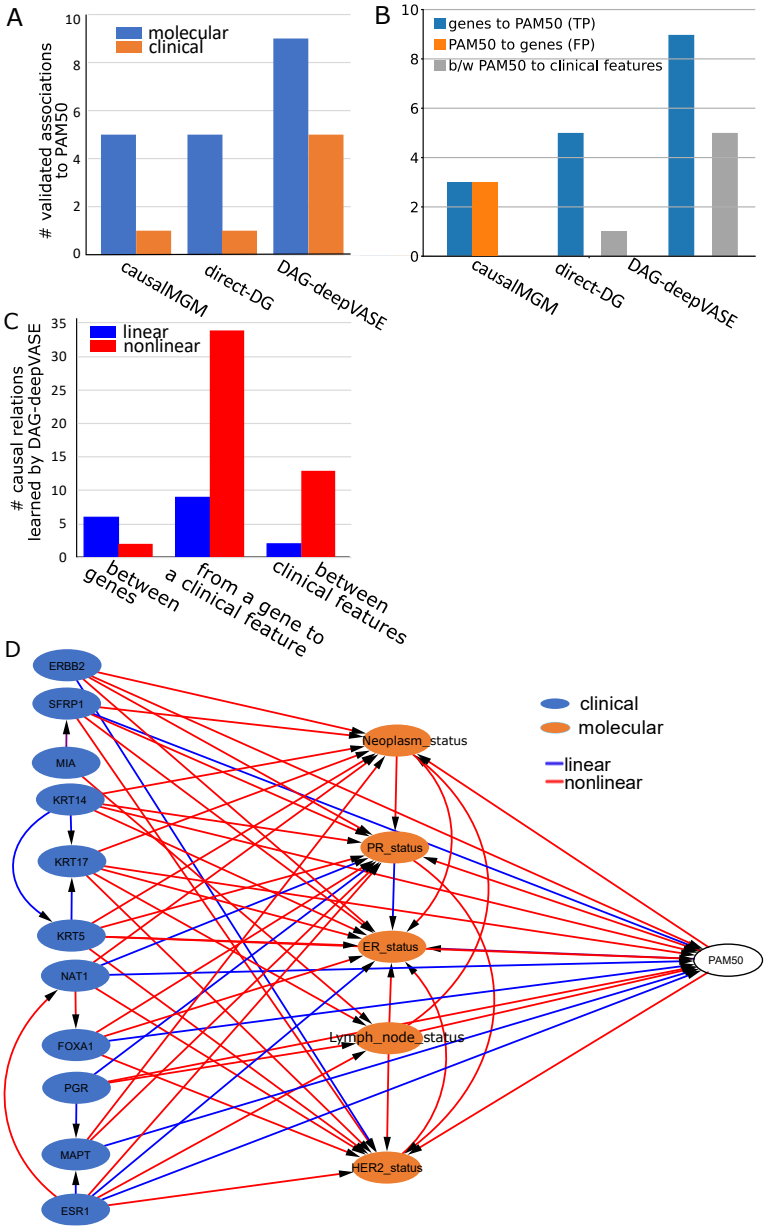
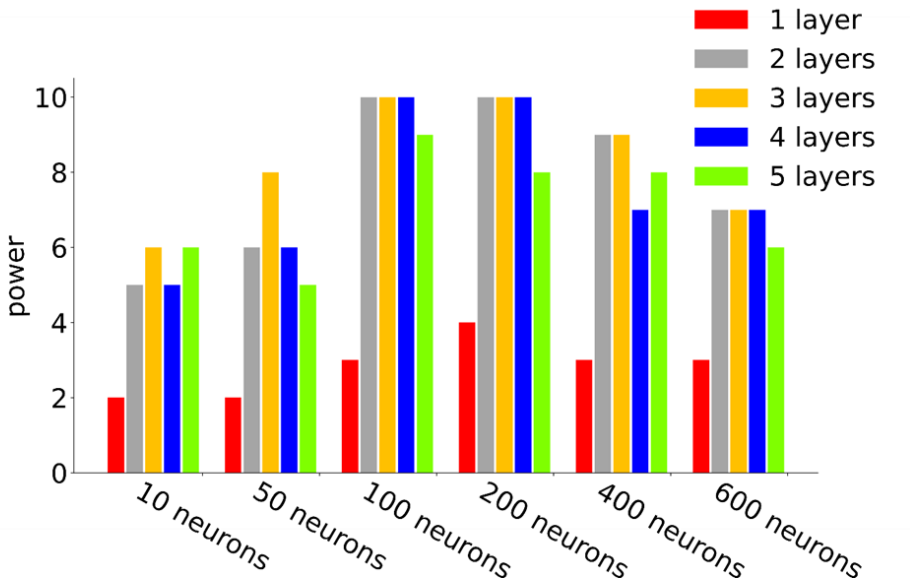


Figure 3.5. DAG-deepVASE on TCGA breast cancer data. (A) Number of validated associations from molecular (blue) and clinical (orange) variables to PAM50 identified by causalMGM, linear-DG, and DAG-deepVASE. (B) Number of causalities identified by causalMGM, linear-DG, and DAG-deepVASE. (C) Number of linear and nonlinear causalities DAG-deepVASE learned between two of the 10 genes, between a gene and a clinical variable, or between two of the 6 clinical variables. (D) Causalities inferred by DAG-

deepVASE over 10 molecular variables, 5 clinical variables, and the PAM50 status as linear (purple) and nonlinear (red) by DAG-deepVASE. 'person_neoplasm_cancer_status' refers to the state or condition of an individual's neoplasm. 'PR_status', 'ER_status', and 'HER2_status' refer to the status of progesterone receptor, estrogen receptor, and human epidermal growth factor 2 receptor in the tumor sample.



Supplemental Figure 8. Average number, and standard error (error bar) of DAG-deepVASE for 10 true associations generated in the complete-nonlinear scenario for 1,000 samples with 190 false associations. To evaluate the model sufficiency, DAG-deepVASE was implemented with various numbers of neuron layers (1~5 layers) and various numbers of neurons (10, 50, 100, 200, 400, and 600 neurons) in each layer.

3.5 Conclusion

We developed the first method, DAG-deepVASE that explicitly learns both linear and nonlinear causal relationships in complex biological systems in high-dimensional molecular and clinical data. In complex biological systems, multiple regulatory layers, e.g., transcriptome and methylation layers, extensively interact [52], [98], [100], [194] and render variable interactions highly nonlinear. In the simulated data of diverse scenarios and biological data of various contexts

(pediatric sepsis, TCGA breast cancer, BMI with nutrients and gut bacteria), DAG-deepVASE consistently outperforms existing methods in identifying known and new nonlinear causal relations. In the first step to identify associations, while DAG-deepVASE identifies all the linear associations that are identified by causalMGM and linear-DG, the method identifies non-linear associations through DNN, which shows the power ranging from 87 % to 100 % in identifying associations. In the second step to identify causalities from the identified associations, DAG-deepVASE infers causalities with a high accuracy (ranging from 88 % to 100 %) while causalMGM learned bidirectional causalities on most of the associations and linear-DG learned only a small number of identified associations correctly. The reason why causalMGM learned bidirectional causalities in our analyses is that it returns a bidirectional causality between variables mediated or confounded by latent variables [8], which very likely exist in molecular and clinical data sets. In contrast, the second step in DAG-deepVASE imposes a model on input variables (x_j) instead of on the conditional distribution of the association (distribution of $x_j|x_i$). This imposition guarantees to identify nonlinear associations even when the model for the association is mis-specified due to absence of latent mediating variables.

To explicitly learn nonlinear associations, DAG-deepVASE leverages a DNN approach differently from other DNN-based causal inference methods by explicitly modelling nonlinearity in individual variable pairs. Previous DNN approaches have been proposed mainly to navigate the intractable search space for the optimal DAG. Although the studies showed that their local optimal DAGs are often comparable to the global ones obtained through expensive combinatorial search, these methods can also return only a stationary-point solution rather than the global optimum. For example, in our simulation experiments, NOTEARS and DAG-GNN showed generally less than 50% of power compared to DAG-deepVASE. However, our experiments also suggest that the

approach of using DNN for navigating the search space may perform better when more samples are collected to construct a more comprehensive search space. For example, the DNN methods identified generally more true causalities when more samples are input (**Figure. 3.2C, 3.2D** and **Supplemental Figure, 5C, D**), though the improvement seems to come at the expense of high false positives (**Figure. 3.2E** and **Supplemental Figure. 5E**).

Another advantage of DAG-deepVASE is the knockoff framework to estimate effect size for nonlinear associations that prioritizes causal relations over simple correlation based on the exchangeability property of the knockoff framework (see Method). The estimated effect size is significantly larger for validated causal relationships than for non-validated ones in both the BMI data (P-value=0.02, Supplemental Figure. 6**P**) and the breast cancer data (P-value=0.03, **Supplemental Figure. 7D**). Based on the rationale that the effect sizes of validated associations are more apparent and thus stronger, these results suggest that the effect sizes estimated by DAG-deepVASE make sense, and these can facilitate translatable findings of the causal relationships by selecting strong causal relations, either linear or nonlinear, to test in downstream experiments or clinical trials. However, care needs to be taken in interpreting the nonlinear effect size as it does not indicate the strength or the direction of causal relations.

DAG-deepVASE enables a further translatable understanding of complex diseases by putting linear and nonlinear associations together. In the subnetwork of pediatric sepsis data presented above, IFN-γ and TNFα are connected through linear and nonlinear associations (gray nodes in **Figure. 3.3B**). Mouse experiments showed that the interaction between IFN-γ and TNFα triggers inflammatory cell death, tissue damage, and mortality in acute immune diseases characterized by “cytokine storm” including lipopolysaccharide (LPS)-mediated sepsis [195]. While it is difficult to identify multiple cytokines involved in the complex interactions, DAG-

deepVASE could identify the interactions between IFN-γ and TNFα via multiple associations of both linear and nonlinear ones, including MCP-1/CCL2. Since MCP-1/CCL2 shows a protective role in a similar mouse model (a polymicrobial sepsis model with LPS) [196], DAG-deepVASE suggests a therapeutic potential to the detrimental interaction between IFN-γ and TNFα.

Despite the clear advantages, DAG-deepVASE has some limitations in improving the clinical relevance of the findings. The first is that DAG-deepVASE cannot take nonordinal categorical variables and take only continuous and ordinal categorical variables that approximately follow Gaussian distribution since model-X knockoff assumes Gaussian distribution. In this project, this condition did not pose any problem as all variables of our interest were either continuous or ordinal categorical. However, in the future, we will generate the knockoff variables for nonordinal categorical variables based on a regression model for nonordinal categorical variables [40]. Second, while DAG-deepVASE can estimate the effect size, it does not estimate statistical significance. Thus, to identify significant causal relations in the future, we will estimate statistical significance of the likelihood ratio test we derived to determine the causal direction in (1). Third, as with other methods of learning causalities from observational data, the validity of the learned causalities depends on how well the data comply with the three causal assumptions: Markov, faithfulness, and sufficiency. However, biological data could violate these assumptions and weaken the applicability of the inference results. For example, since multiple biological layers, such as genomic, transcriptomic, and epigenetic layers, often interact to render a phenotype in humans, confounders can occur in any of the layers. However, it is not always feasible to measure all variables from all the layers due to technical and practical reasons, indicating that the causal sufficiency assumption of no latent confounder would be hardly met for biological data. Thus, it

is necessary to conduct further experiments or clinical trials to validate the causal relationship learned through DAG-deepVASE.

In summary, we developed DAG-deepVASE, which learns causal relationships in complex biological systems. DAG-deepVASE is the first method that uses a DNN approach to identify linear and nonlinear associations and learn their causal directions. DAG-deepVASE outperforms existing methods, causalMGM, and linear-DG, in identifying known causal relations in various simulation scenarios and molecular and clinical data sets. In addition to known causalities, DAG-deepVASE identifies novel complex pathobiological interactions involving nonlinear causal relations, which is not possible using other methods. By applying the knockoff framework to DNN, DAG-deepVASE estimates effect size for nonlinear associations that prioritizes causal relations, which allows to prioritize future clinical and experimental validations. With these advantages, the application of DAG-deepVASE can help identify driver genes and therapeutic agents in biomedical studies and clinical trials.

4.0 Project 3 - Deep Neural Network Jointly Learning Gene Expression and Biological Condition Information Identifies Cell Subtypes Nonlinearly Linked to the Biological Condition

4.1 Summary

With recent advent of single-cell level biological knowledge, interests grow in identifying the cell states or subtypes that are not only representative of the molecular behavior in terms of gene expression, but also linked to the biological condition of interest, such as disease samples versus normal samples. Since no method has been developed to identify such condition-specific cell subtypes, existing approaches undertake a two-step process where cell clusters of homogenous molecular behavior are first identified based on gene expression information and some of those clusters that are enriched in the biological condition of interest are further selected as condition-specific cell subtypes. However, this approach can lead to suboptimal solutions due to three limitations: 1) it does not consider the impact of one criterion on another, 2) it disregards the dimensional differences in the criteria, and 3) the optimizations rely on linear modeling. To address the limitations and accurately identify such condition-specific cell subtypes, we present scDeepJointClust, the first method that addresses the limitations by jointly training on both types of information in a deep neural network (DNN) approach. Using scDeepJointClust on simulation data of multiple scenarios and biological data of various contexts, we demonstrated the superiority of scDeepJointClust over existing methods in terms of sensitivity and specificity, holding significant promise for advancing our understanding of cellular states and their implications in complex biological systems.

4.2 Introduction

The expansion of single-cell measurements, e.g., single-cell RNA-Seq (scRNA-Seq) data, allows researchers to identify cellular states or subtypes that exhibit a homogeneous molecular behavior in terms of gene expression. Further, with the recent progress in our understanding of cellular states, there is growing interest in identifying cell states or subgroups that serve as representatives of gene expression patterns and exhibit enrichment in specific biological conditions. These condition-specific cell states encompass various contexts, including distinguishing disease samples from normal samples, identifying specific stages among multiple developmental stages, and differentiating experimental intervention samples from control experiments. The ability to accurately identify and characterize such condition-specific cell states holds immense potential for advancing our understanding of complex biological systems and their implications in disease mechanisms and therapeutic interventions [197], [198]. The significance of identifying condition-specific cell subgroups extends to both biological and clinical domains. From a biological standpoint, cells of the same type exhibit distinct states based on their developmental stage, function, and responses to external stimuli [199]–[201]. Thus, the identification of diverse states in the same cell type aids researchers in comprehending the intricate mechanisms driving tissue development, immune responses, and disease progression. From a clinical perspective, understanding if specific cell states correlate with improved clinical outcomes under particular treatment regimens holds importance since such knowledge can pave the way for

targeted therapies that have the potential to significantly enhance disease treatment and management [202].

Despite the significance of identifying condition-specific cell subtypes, existing methods have not explicitly utilized biological condition information for this purpose. Instead, a common approach involves a two-step process that employs different methods in each step. In the first step, cell clusters are defined based mostly on the molecular behavior represented in the gene expression information. To achieve this, several methods embed the gene expression information of cells into a graph structure, such as a k-nearest neighbor graph, and detect dense regions in the graph through community detection algorithms. These clusters are then classified into cell types. In the second step, the pre-defined cell clusters are further examined to identify cell subtypes that are enriched in the biological condition of interest compared to other condition(s) (e.g., tumor samples versus. normal), often referred to as a differential abundance test. For example, on the scRNA-seq data of immune cells from 35 non-small cell lung cancer (NSCLC) samples and 29 matched healthy non-involved samples [203], a recent study pre-defined 30 cell clusters of multiple cell types (e.g., B, Mast, macrophage, natural killer (NK) cells) based on the gene expression information. Then, by quantifying and comparing their abundance in the tumor versus healthy samples, specific clusters were identified as cell states correlated with an enhanced response to immunotherapy, such as PDCD1+CXCL13+ activated T cells, IgG+ plasma cells, and SPP1+ macrophages. Despite these findings, the current approach lacks explicit integration of biological condition information into the identification of condition-specific cell subtypes.

Indeed, the two steps can be reversed in the analysis pipeline, offering an alternative approach to identifying condition-specific cell subtypes. Specifically, one could first conduct a differential abundance test without the definition of cell clusters and then further perform

clustering based on the test result. To implement this approach, practitioners can use a method like Milo. Milo was recently developed to allow users to perform differential abundance analysis by the unit of neighbors, sets of random cells Milo identified to be similar in the gene expression profile. Since the neighbors redundantly represent mixture of cells and not all the cells are sampled in the neighborhood representation, Milo cannot be directly used to define cell clusters. However, practitioners can still overlay the differential abundance result with predefined cell clusters on the same dimensionality-reduced feature space and visualize how the predefined cell clusters align with the abundance results, leading to potential refinements that better reflect the cell states linked to the biological condition. Since Milo can also cluster cells using k-nearest neighbor algorithm and the corresponding latent space without explicitly using the biological condition information, it presents a suitable choice for our comparison experiment.

In both of the two-step approaches described above, we identify three limitations that can hinder the accurate identification of condition-specific cell states or subtypes. First, the approaches do not consider potential interactions between the two criteria, gene expression and the biological condition information. Since samples of a particular biological condition would render distinct biological functions represented with distinct molecular behavior, it is expected that the distinct molecular behavior is represented in the level of gene expression. Due to this interaction, optimizing one criterion after another would not be able to model the interaction and thus be a less integrated and holistic approach compared to training on the two criteria simultaneously, or jointly training on the criteria that can balance the two criteria in a single optimization process. Secondly, when identifying cell states based on both gene expression information and the biological condition, it is essential to explicitly control the weight of the two criteria. While the gene expression information typically represents tens of thousands of genes, we are interested in a

particular type of biological condition at any given time. Thus, when training a model on the gene expression information and the biological information separately, the biological condition's weight and influence can be effectively diluted and diminished during the optimization process. Thirdly, it is essential to learn the complex relationship in how the cell states are defined with respect to gene expression and the biological condition information. Cells undergo a series of differentiation events that lead to the formation of distinct cell types with specialized functions, rendering a unique set of molecular characteristics that nonlinearly determine their identity and functional properties. However, most existing methods do not fully consider the nonlinear relationship of cell identity. For example, in case of Milo, which relies on a simple k-nearest neighbor (KNN) data structure, the ability to uncover the complex relationship among cells is limited. As a result, the method may not fully exploit the richness of the data and might overlook critical associations between cell states and the biological condition of interest.

In this project, we present scDeepJointClust as a solution to address the aforementioned limitations. Firstly, to enable simultaneous training on both gene expression and the biological condition information, scDeepJointClust adopts a joint-learning approach where the model is simultaneously trained on both information. Second, in order to effectively control the weights assigned to the gene expression information and the biological condition in the joint learning process, scDeepJointClust takes the molecular status as input to represent the whole gene expression information. Given the considerable sophistication of existing methods in capturing the molecular status represented in gene expression information, scDeepJointClust adopts an approach where it can take the result of such a method as an input, along with the biological condition information. This design allows practitioners to effectively control the weights between the gene expression information and the biological condition, empowering them with the flexibility to

adjust the contribution of each factor. This flexibility further allows to update the identification when future methods emerge to represent the molecular behavior more accurately from gene expression information. Since methods are being developed to improve performance of cell clustering based on gene expression, this computational adaptability allows scDeepJointClust to directly import the improvement to more accurately identify condition-specific cell states or subtypes. Thirdly, to learn the complex and nonlinear relationship in how cell states are defined with the input data, scDeepJointClust utilizes the DNN method that can model the nonlinearity with a number of neuron layers.

scDeepJointClust represents the first attempt, to the best of our knowledge, to explicitly identify cell states that are not only representative of gene expression but also associated with biological conditions using a Deep Neural Network (DNN) model in a joint learning framework. Previously, DNN joint learning approaches have demonstrated successful applications across diverse scientific domains, effectively integrating multiple sources of information to solve complex problems. For instance, in the context of speech recognition where DNN models have been extensively utilized, a multi-feature and multi-task DNN method learns multiple acoustic features to successfully enhance language recognition performance [204]. Similarly, to classify images, a DNN method was proposed to consider both class label information and local spatial details, exhibiting remarkable accuracy on various benchmark datasets when compared to baseline methods [205]. Remarkably, the DNN method for image classification addresses a problem structure akin to ours where the class label and local information are replaced by the biological information and gene expression information. Due to this similarity, it underscores the potential of the DNN joint learning approach in effectively solving our specific problem. DNN approaches have been used in the context of single-cell RNA-Seq data, albeit without the utilization of joint

learning. For instance, a DNN model was proposed to correctly solve cell-type related problems, such as identifying new cell types and states, by integrating pathway knowledge [206] as prior knowledge. Also, a recent work selects genes whose expression pattern are shared by the cells of the same type by reducing representation in the output layer of denoising autoencoder [207] with neural approximator (DAWN) and pairing this reduced representation with the model-based EM clustering.

Using scDeepJointClust on simulation data of multiple scenarios and real biological data of different contexts, such as patients with advanced melanoma and non-small cell lung cancer undergoing ICB), we demonstrate the advantage of using scDeepJointClust over existing methods in terms of sensitivity and specificity. Altogether, we develop a DNN-based joint-learning method that simultaneously optimizes the information of gene expression and the biological condition to successfully identify cellular states linked to a biological condition with the highest sensitivity and specificity.

4.3 Materials and Methods

4.3.1 Model Setting

The proposed method consists of an input layer, a dropout layer, M hidden layers, a cell type classification output layer, a responder-nonresponder classification output layer. This model takes one input data and two labeled data. The input data is a single-cell level gene expression

matrix X consisting of N cells (in rows) and D genes (in columns). The first labeled data is cell type matrix Y^t in which each row $y_i^t = (t_1, \ldots, t_J)$, where J is the number of cell clusters defined using gene expression information, where $t_j = 0$ if cell i belongs to cluster j and 0 otherwise. The second labeled data is cell origin label matrix Y^r in which each row $y_i^r = (r_1, \ldots, r_K)$, where K is the number of biological conditions, where $r_k = 0$ if cell i belongs to condition k and 0 otherwise. The input layer passes on its output to the dropout layer. The dropout layer randomly sets 5% of its neuron units to 0 at each step of the training procedure. As a result, only relevant genes would be selected as parts of the final representation layer. For each of the M hidden layers m, we use Glorot normal initializer (Xavier normal initializer) for initializing the layer weights and Rectified Linear Units (ReLU) as its activation function. The output of m can be described as follows:

$$o_m = ReLU(W_m o_{m-1} + b_m)$$

where W_m is the weight matrix, o_{m-1} is the output of the previous layer, and b_m is the bias term for this layer.

As cell types are perceived as classes, we compute the loss between the cell type output layer's predictions and the true labels based on categorical cross-entropy and use softmax as the activation function for the cell type output layer. The output of the cell type output layer can be described as follows:

$$\bar{y}^t = softmax(W_t o_m + b_t)$$

where W_t is the weight matrix, o_m is the output of the last hidden layer, and b_t is the bias term for the layer.

If the biological conditions are multiple, we use the same softmax function to optimize. However, since biological conditions are often binary (e.g., case vs. control or responder vs. non-responder), we provide an option of it being binary and compute the loss between the responder-

nonresponder output layer's predictions and the true labels based on cross-entropy and use sigmoid as the activation function for the output layer. The output of this layer can be described as follows:

$$\bar{y}^r = sigmoid(W_r o_m + b_r)$$

where W_r is the weight matrix, o_m is the output of the last hidden layer, and b_r is the bias term for the layer.

Objective function and optimization

Since our method is a joint learning algorithm, the objective function includes two loss functions, a classification loss from the cell type predictions, a classification loss from the biological condition predictions. The cell type prediction loss is calculated as follows:

$$L_t = -\sum_{n=1}^{N}\sum_{j=1}^{J} y_{jn}^t log\bar{y}_{jn}^t$$

where y_n^t and $\bar{y}_n^t$ are two vectors in a one-hot representation, y_{jn}^t is the truth value (0 or 1) of jth element in the one-hot vector y_n^t, and $\bar{y}_{jn}^t$ is the predicted probability of x_n being categorized as jth cell type.

The cell origin prediction is calculated as follows:

$$L_r = -\sum_{n=1}^{N} y_n^r log\bar{y}_n^r + (1 - y_n^r)\log\,(1 - \bar{y}_n^r)$$

where y_n^r is the truth value of x_n being a responder and $\bar{y}_n^r$ is the predicted probability of x_n being a responder.

With these two losses, the objective function for the proposed method can be described as follows:

$$min - \lambda_t \sum_{n=1}^{N}\sum_{j=1}^{J} y_{jn}^t log\bar{y}_{jn}^t - \lambda_r \sum_{n=1}^{N} y_n^r log\bar{y}_n^r + (1 - y_n^r)\log(1 - \bar{y}_n^r)$$

where λ_t is a user-defined value for controlling how much emphasis should be put on cell type information, λ_r is a user-defined value for controlling how much emphasis should be put on responder-responder information.

To minimize this objective function, we use the Adam optimization algorithm, which is a stochastic gradient descent method, to retrieve the optimal network parameters Θ.

4.3.2 Clustering Module

For this work, we utilized the K-Means clustering algorithm [208] in the clustering module, and the implementation of the algorithm is provided and maintained by scikit-learn [209].

K-Mean clustering algorithm aims to partition the input data X with N observations into K clusters. The algorithm works in the following fashion:

1) Randomly selects K centroids as the beginning points for each cluster.
2) For each data point X_n, calculate the distance between X_n and centroid C_k using a distance metric, and assign X_n to its closest cluster centroid C_k.
3) For each cluster, calculate the average of all the data points in this cluster and re-initialize its centroid based on the average.
4) Keep repeating steps 2 and 3 until there is no change in the assignments of data points to clusters, meaning the centroids have stabilized.

4.3.3 Simulation Data

We used the R package dyntoy [210] to generate simulated single-cell datasets with discrete clusters. For each simulated dataset, we generated 10 discrete clusters. In each of these clusters, we assign 90% of the cells to one of two simulated biological conditions (C1 and C2) and the rest 10% of the cells to the other condition. If 5 of these 10 clusters are dominated by condition C1 (90% of the cells belong to C1), then the other 5 clusters were populated with the other

condition C2 (90% of the cells belong to C2), and vice versa. This simulation procedure was introduced by Milo [211].

4.3.4 Non-Small Cell Lung Cancer (NSCLC) Single-Cell RNA Sequencing (scRNA-seq) Data

NSCLC CITEseq dataset was presented in Leader et al [203]. We downloaded the NSCLC metadata from their GitHub repository: https://github.com/effiken/Leader_et_al. The scRNA-seq data pre-processing workflow in this work, e.g., the selection and filtration of cells based on QC metrics, data normalization and scaling, and the detection of highly variable features, was performed using the R package Seurat [212] Seurat, an R toolkit, is widely used in the field of computational biology to analyze scRNA-seq data.

4.3.5 Benchmarked Clustering Methods

We evaluated our method against three existing clustering methods. This section provides details on what packages were used and how they were run.

- Louvain [213]: Louvain algorithm is a popular hierarchical clustering method used to identify communities within complex biological networks. This algorithm calculates a modularity score for each community to maximize the detection of communities. The modularity score is typically used to evaluate how well nodes are assigned to communities. The Louvain implementation used in this work is also supported by the Python package Scanpy [214]. PCA was also computed before Louvain was performed.
- Leiden [215]: Leiden is also a hierarchical clustering algorithm, which is based on Louvain. The algorithm has been modified to address the issue of poorly connected communities. This is done by periodically breaking down the communities into smaller, more well-connected ones. All the Leiden clustering runs were performed using the Python package Scanpy [214], which is a scalable toolkit for analyzing single-cell gene

expression data. We performed principal component analysis (PCA) on the data before running Leiden.

- Milo: Milo is developed to perform differential abundance testing by assigning cells to partially overlapping neighborhoods on a k-nearest neighbor graph. We used the Python implementation of Milo algorithm (https://github.com/emdann/milopy) [211]. As Milo requires a k-nearest neighbor (KNN) graph before performing its downstream analysis, we used the KNN implementation from the Python package Scanpy [214] to build such graphs.

4.3.6 Single-Cell RNA Sequencing Annotation

The cell type annotation was performed using an R package called 'SingleR' [216], a novel computational method for performing unbiased cell type annotation on scRNA-seq data. SingleR annotates each cell by leveraging a reference transcriptomic dataset of pure cell types. The reference transcriptomic dataset used in this work is generated and supplied by Blueprint and ENCODE [203], [210]–[212], [214], [217].

4.4 Results

4.4.1 Modeling Cell States by Jointly Training on Gene Expression and the Biological Condition Information

To accurately identify cell states linked to a biological condition (e.g., tumor vs. normal samples), we present scDeepJointClust (**Figure. 4.1A**) to address the limitations in the current

two-step approaches. To address the first limitation and jointly train on both gene expression and the biological condition information, scDeepJointClust simultaneously optimizes two loss functions, L_t for the gene-expression information and L_z for the biological condition information of the cells (see Methods). To simultaneously address the second limitation and control the weight of the gene expression information vs. that of the biological condition information, scDeepJointClust sets L_t with the cluster information generated by a method of user's choice. By optimizing L_t this way, scDeepJointClust will first embed the cluster structure onto the scDeepJointClust model. Furthermore, by optimizing L_z, we can enhance the identification of cell states that hold significant importance in a biological condition. As illustrated in **Figure. 4.1B**, cell clusters often exhibit overlapping gene expression patterns, such as cluster 3 overlapping with cluster 2 in the illustration. If the clusters convey valuable biological signals associated with a specific biological condition under investigation, they are expected to exhibit certain levels of enrichment or depletion in terms of the biological condition information. For example, in **Figure. 4.1C**, cluster 1 and 3 show an enrichment of condition A vs. condition B while cluster 2 shows an enrichment of condition B vs. condition A. If condition A represents a disease state (e.g., tumor) and condition B represents a control state (e.g., normal), then clusters 1 and 3 would represent the cells associated with biological processes related to the tumors. To solve the third limitation and accurately model the complex relationship in the cell state assignment, scDeepJointClust uses a DNN component to encapsulate the complex and nonlinear relationships using multiple layers of nonlinear activators called neurons (**Figure. 4.1A**). Specifically, the neurons in the first layer learn to extract low-level features from the gene expression data, while neurons in the subsequent layers learn to extract higher-level features. Each layer acts as a nonlinear transformation of the input

data, making the model more expressive and capable of capturing the complex relationships between the information and the cell states.

Concretely, scDeepJointClust employs several DNN techniques to accurately pick up signal from single-cell data. First, scDeepJointClust utilizes dropout to prevent overfitting (see Methods). Overfitting is a serious problem in single-cell data analysis because single-cell data is high-dimensional and noisy [218]–[221], and it is easy to fit a model that captures the noise additional to the underlying biological signal, leading to false discoveries and misinterpretation of the results [222]–[225]. By dropping out neurons, scDeepJointClust is forced to learn redundant representations and is less likely to rely on a few neurons to make decisions, encouraging a more robust and less overfitting estimation. Second, scDeepJointClust allows the learning on the biological condition of multiple values by adaptively designing the cell type output layer with either softmax (for multiple values) and sigmoid (for binary values). This adaptable design is crucial because it allows for the analysis of single-cell data in both two-group and multi-group scenarios. As an example of the two-group scenarios, normal samples are frequently analyzed alongside tumor samples to gain insight into tumor biological processes relative to the corresponding normal tissue state. An example of a multi-group comparison is the analysis of embryonic brain development where researchers can examine the gene expression profiles of brain tissue samples collected at multiple developmental stages, such as early embryonic, mid-embryonic, and late embryonic stages. By analyzing the multi-group data collectively, scDeepJointClust can identify cell clusters that exhibit enrichment for any of the multiple biological conditions, thereby effectively revealing crucial cell subtypes that are characteristic in the developmental stages.

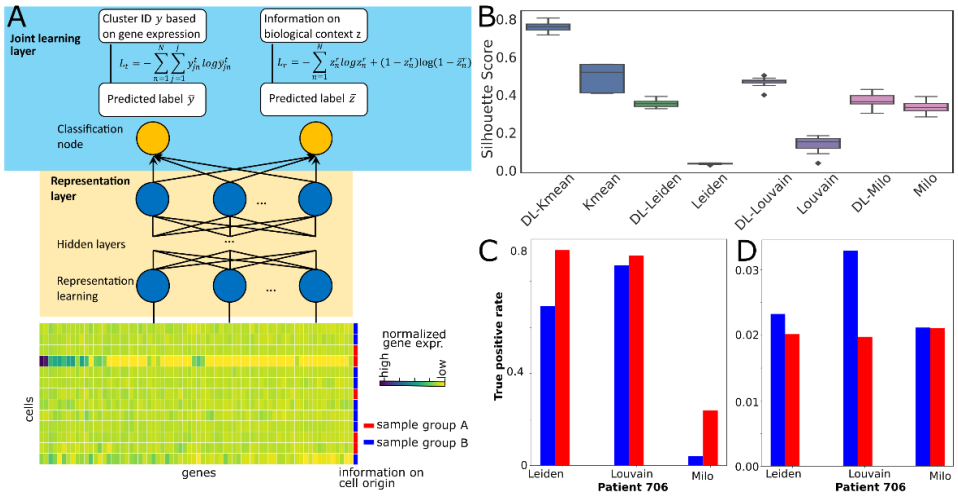


Figure 4.1. Overview of scDeepJointClust. (A) An input matrix of gene expression and the biological condition information (columns) over a set of cells (rows) is fed into the DNN component. Then, the representation layer will transfer the training result to two output layers so the model can be optimized with two criteria L_t and L_r. (B) An example 3 cell clusters (cluster 1 in gray, cluster 2 in green, and cluster 3 in yellow) on UMAP using the gene expression information. (C) The cells in the example clusters are presented with two biological conditions, A or B, where each cluster is characterized with one of either condition, cluster 1 and 3 with condition A and cluster 2 with condition B.

4.4.2 scDeepJointClust Refines Pre-Defined Cell Clustering Results With Condition Information

To evaluate the performance of scDeepJointClust in the presence of gene expression information and the biological condition information, we simulated 10 clusters of 10,000 cells each cluster characterized by both types of information (see **Materials and Methods**). The gene expression information was simulated to reflect varying distances between cell clusters, representing different degrees of similarity among them, as is the case in biological data. The biological condition was simulated to exhibit differential enrichment degrees across the cell types.

For example, we assigned a higher proportion (e.g., 80%) of a biological condition to several clusters, while allocating a lower proportion (e.g., 20%) of the biological condition to the remaining clusters. With this experimental design, we conducted two scenarios of simulation experiments. In the first scenario, we simulated two distinct biological conditions, along with gene expression profiles, to replicate case-control sequencing experiments (tumor vs. normal), where samples from the case group are sequenced alongside control group samples and analyzed in relation to each other (**Figure. 4.2A**). On the simulated gene expression data, we ran established clustering methods that take only gene expression information (K-means, Leiden, Louvain, and Milo). Then, on each of the clustering results, we ran scDeepJointClust with the biological condition information. After repeating this experiment 100 times, we compared the clustering result to the ground truth definition of cell clusters in terms of Silouette score, demonstrating that scDeepJointClust always outperforms the clustering results of the other methods (**Figure. 4.2B**). Specifically, scDeepJointClust demonstrates over a twofold enhancement compared to K-means-, Leiden-, and Louvain-based clustering outcomes, while it also markedly improves the clustering result of Milo.

Further, we simulated another scenario where three biological conditions, instead of two, are spread over 10 clusters of cells with varying distances from each other (**Figure. 4.2C**). scDeepJointClust outperforms the clustering results of the established methods almost equally as the case of two biological conditions with over a twofold improvement. Overall, the superior performance of scDeepJointClust illustrates how it takes advantage of the biological condition information to better identify cell types that are enriched with a specific biological condition.

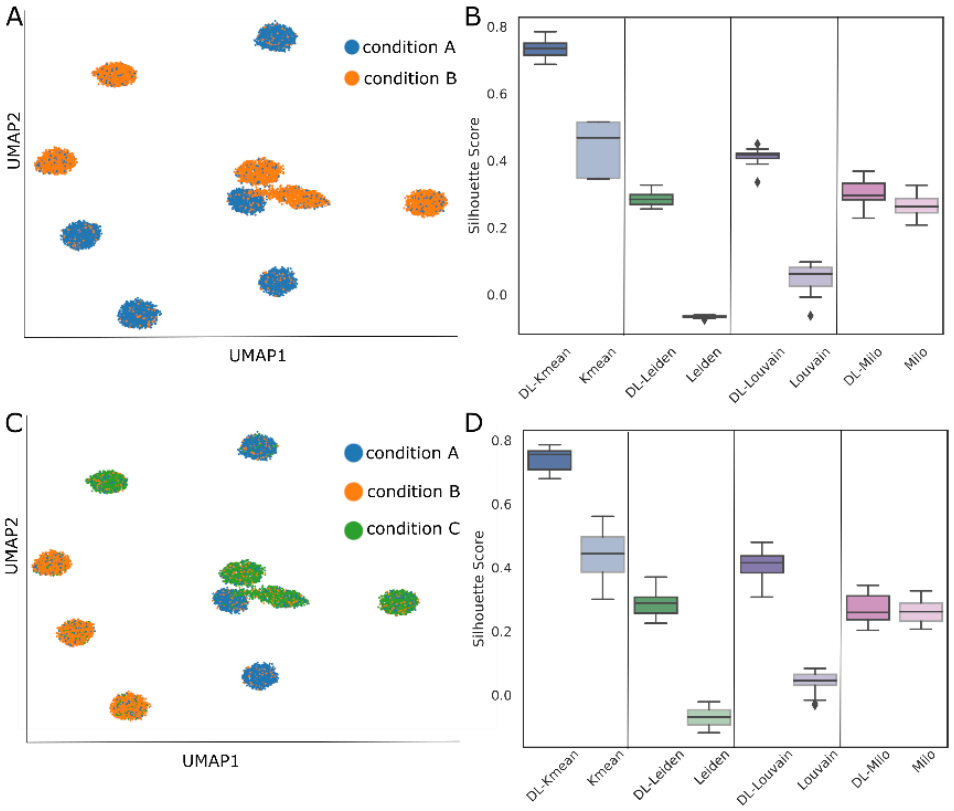


Figure 4.2. Performance assessment using simulation data. (A) UMAP visualization of 10 simulated cell clusters using the gene expression information. On the visualization, the cells are colored by two simulated biological conditions, A or B. (B) Methods' performance in silhouette score in 100 random trials for each experiment. While Kmeans, Leiden, Louvain, and Milo represent the clustering result of the methods in terms of Silouette score, DL-Kmeans, DL-Leiden, DL-Louvain, and DL-Milo represent the result of the refinement brought by scDeepJointClust. (C) UMAP visualization of 10 simulated cell clusters using the gene expression information. On the visualization, the cells are colored by three simulated biological conditions, A, B or C. (D) Methods' performance in silhouette score in 100 random trials for each experiment. While Kmeans, Leiden, Louvain, and Milo represent the clustering result of the methods in terms of Silouette score, DL-Kmeans, DL-Leiden, DL-Louvain, and DL-Milo represent the result of the refinement brought by scDeepJointClust.

4.4.3 scDeepJointClust Embeds Pre-Defined Cell Clustering Results in A Deep Neural Network Model

One of the key features of scDeepJointClust is its utilization of a deep neural network (DNN) structure to embed a gene-expression-based clustering outcome (L_t). Ensuring the high quality of this embedding is essential because the joint training with the biological condition information (L_z) will rely on this embedding. To test quality of the embedding, we trained scDeepJointClust's DNN structure only with a gene-expression-based clustering outcome and evaluate if the training can effectively incorporate the clustering outcome by separating the clusters. For this, we downloaded a single cell RNA-Seq data set of 16,291 immune cells from 48 tumor samples of melanoma patients treated with immune checkpoint therapy (Pembrolizumab, anti-PD1) [226]. From the cells, the original study identified 11 cell types based on a list of known marker genes and a manual review process. The identified cell types include B, Plasma, Monocyte/Macrophages, Dendritic cells, etc. We visualized the cell types on a t-SNE plot based on gene expression data and labeled them by the original clustering result (**Figure. 4.3A**). Despite a rough separation observed on the t-SNE plot, the cell types are not perfectly separated from each other by the gene expression. The lack of clear separation in the gene-expression-based projection implicates a challenge in deriving the clusters solely based on gene expression. However, when we specifically trained scDeepJointClust based on the clustering result in the original paper (**Figure. 4.3B**), the representation layer of scDeepJointClust (the dark blue layer in **Figure. 4.1A**) further separates the clusters so it can effectively facilitate the subsequent joint learning processes with the biological condition information.

To ensure generalizability of this model behavior, we downloaded another single cell RNA-Seq data set of 361,929 cells from 35 early-stage NSCLC lesions. By running an

unsupervised batch-aware clustering method [227] on the data, they derived 60 cell clusters representing diverse sub cell types of immune cells such as Natural Killer (NK), T, Mononuclear Phagocyte (MNP), plasmacytoid Dendritic Cell (pDC), B, plasma, and MAST cells. In line with the melanoma data, the t-SNE visualization of the cells solely based on the gene expression information demonstrates not clear separations among the cell types (**Figure. 4.3C**). However, upon training scDeepJointClust only with the gene-expression-based clustering outcome, the representation layer effectively segregates the clusters (**Figure. 4.3D**). With the melanoma data and the NSCLC data using different clustering methods on the gene expression information, our results consistently showed the lack of separation in the cell type clusters on the original gene-expression feature space and an improved separation of them when trained on scDeepJointClust's DNN model, demonstrating that the DNN model effectively facilitates the subsequent joint learning processes which will be with the biological condition information.

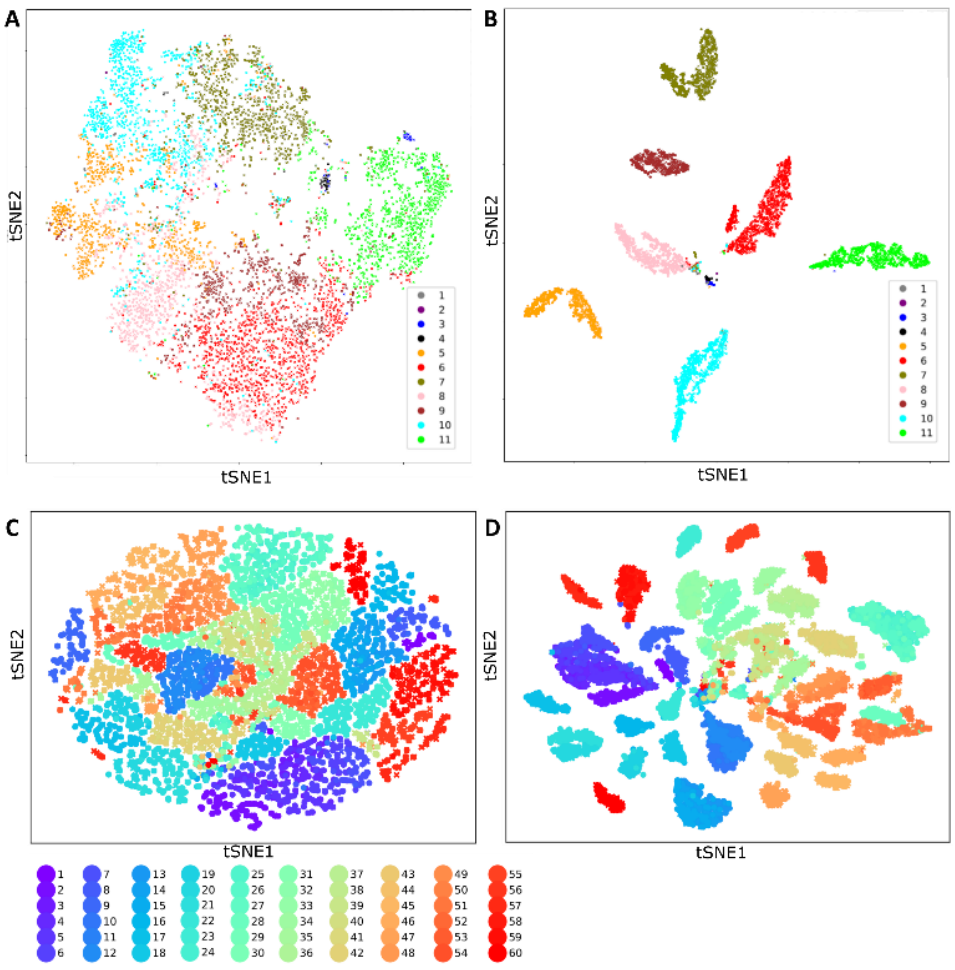


Figure 4.3. Evaluation of embedding performance. (A) tSNE of the melanoma single cell data using the gene expression profile. (B) tSNE of the melanoma single cell data after training on L_t in accordance with 11 input clustering result. (C) tSNE of the NSCLC single cell data using the gene expression profile. (B) tSNE of the NSCLC single cell data after training on L_t in accordance with 60 input clustering result.

4.4.4 scDeepJointClust Identifies Cell Clusters Correlated With Enhanced Response to Immunotherapy

To evaluate the performance of scDeepJointClust using both gene expression and the biological condition information, we further analyzed the NSCLC tumor samples and patient-

matched healthy (noninvolved) lung samples (nLung) since it is one of few data sets that provides a CITE-Seq data from two experimental conditions (tumor and nLung samples). CITE-Seq data is a combination of gene expression and antibody information where the antibodies represent cell epitopes that play important roles in identifying the true cell types. Thus, we measured true positive and false positive rate (TPR and FPR) in subsequent experiments against the cell cluster IDs published in the original paper that were based on the gene expression and antibody information. Since scDeepJointClust is designed to identify cell types that are enriched with a particular condition, we selected four cell subtypes that showed an enrichment to either tumor samples or nLung samples in the original paper and its replication data set, which are NK, B, T, monocyte and macrophage (momac) cells. NK cells, B cells, and T cells are part of the comprehensive immune defense system, collectively working to detect, eliminate, and remember specific pathogens while maintaining the overall health of the body. In both the original data and a replication data of NSCLC [228] vs. normal samples, NK cells are enriched in nLung samples vs. tumor samples and momac, T, and B cells are enriched in tumor samples vs. nLung samples. For targeted analysis, we specifically chose two patients with higher counts of the cells out of 7 patients in the data, namely 695 and 703. To evaluate how scDeepJointClust refines the result of existing clustering methods in patient 695 and 703, we first clustered the cells based only on the single cell RNA-Seq data using existing clustering methods (Leiden, Louvain, and Milo). After annotating the clusters into cell types using a computational method, singleR, based on the Human Primary Cell Atlas reference dataset [216], we assessed how well the identified NK, B, and T cells match those identified in the original publication that we deemed true. Then, we ran scDeepJointClust on the clustering result with the biological condition information (tumor vs. nLung) and compared the results.

In terms of TPR, scDeepJointClust enhances the result of all the tested methods for all four cell types in both patients with only few exceptions in identifying B and momac cell types (**Figure. 4.4A, 4.4B, Table 4.1**). Furthermore, scDeepJointClust also enhances the FPR in both patients with a single exception in identifying T cell types (**Figure. 4.4C, 4.4D, Table 4.1**). Given the uncertainty in choosing a gene-expression-based clustering algorithm, it is worth emphasizing that the improvement provided by scDeepJointClust remains substantial across all clustering algorithms tested. It is interesting to note that, when there is little gain in either TPR or FPR, scDeepJointClust substantially refines the result in the other criteria, refining the overall results of cell state identification. For example, although scDeepJointClust does not improve TPR of momac identification for Patient 695 from any existing algorithms, scDeepJointClust makes drastic improvements from all the methods in terms of FPR. Similarly, for Patient 695, while scDeepJointClust does not improve false positive rate of B, NK, and T cell identification, it improves true positive rate of the cells.

Table 4-1. Parameter settings for the deep-learning component of DAG-deepVASE.

Patient 695					
TPR for B	Leiden	0.979147	0.963033	-0.01611	-0.016
	Louvain	0.977251	0.82019	-0.15706	-0.157
	Milo	0.033175	0.15327	0.120095	0.12
TPR for NK	Leiden	0.802244	0.957924	0.15568	0.156
	Louvain	0.821879	0.945302	0.123422	0.123
	Milo	0.046283	0.160237	0.113954	0.114
TPR for T	Leiden	0.939153	0.936316	-0.00284	-0.003
	Louvain	0.925702	0.952969	0.027267	0.027
	Milo	0.035044	0.671516	0.636472	0.636
TPR for MoMac	Leiden	1	1	0	0
	Louvain	1	1	0	0
	Milo	1	1	0	0
FPR for B	Leiden	0.002786	0.002502	-0.00028	0
	Louvain	0.003405	0.002743	-0.00066	-0.001

	Milo	0.024529	0.023448	-0.00108	-0.001
FPR for NK	Leiden	0.029759	0.024458	-0.0053	-0.005
	Louvain	0.027385	0.022598	-0.00479	-0.005
	Milo	0.023581	0.022231	-0.00135	-0.001
FPR for T	Leiden	0.020647	0.021292	0.000645	0.001
	Louvain	0.015352	0.014963	-0.00039	0
	Milo	0.013982	0.148728	0.134746	0.135
FPR for MoMac	Leiden	0.156359	0.160935	0.004576	0.005
	Louvain	0.156586	0.134113	-0.02247	-0.022
	Milo	0.961896	0.389561	-0.57233	-0.572
Patient 706					
TPR for B	Leiden	0.998094	0.988325	-0.00977	-0.01
	Louvain	0.997856	0.826123	-0.17173	-0.172
	Milo	0.039552	0.475578	0.436026	0.436
TPR for NK	Leiden	0.693069	0.936634	0.243564	0.244
	Louvain	0.871287	0.912871	0.041584	0.042
	Milo	0.041584	0.239604	0.19802	0.198
TPR for T	Leiden	0.860237	0.950734	0.090496	0.09
	Louvain	0.720881	0.722573	0.001692	0.002
	Milo	0.037254	0.566715	0.529461	0.529
TPR for MoMac	Leiden	1	0.981481	-0.01852	-0.019
	Louvain	1	1	0	0
	Milo	1	0.888889	-0.11111	-0.111
FPR for B	Leiden	0.185023	0.183857	-0.00117	-0.001
	Louvain	0.205181	0.174297	-0.03088	-0.031
	Milo	0.020753	0.020848	9.51E-05	0
FPR for NK	Leiden	0.024738	0.02002	-0.00472	-0.005
	Louvain	0.033178	0.019777	-0.0134	-0.013
	Milo	0.021581	0.020212	-0.00137	-0.001
FPR for T	Leiden	0.026304	0.027393	0.001089	0.001
	Louvain	0.018643	0.017311	-0.00133	-0.001
	Milo	0.018943	0.019873	0.00093	0.001
FPR for MoMac	Leiden	1	0.981481	-0.01852	-0.019
	Louvain	1	1	0	0
	Milo	1	0.888889	-0.11111	-0.111

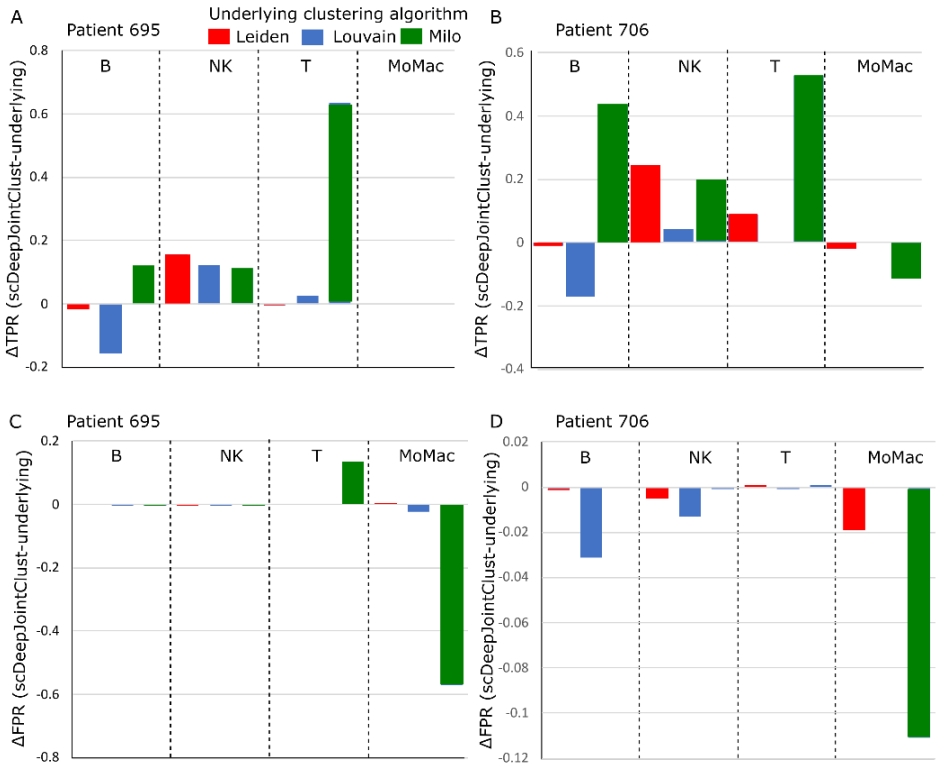


Figure 4.4. Demonstration of scDeepJointCluster in NSCLC data. TPR difference of scDeepJointClust vs. existing methods (Leiden in red, Louvain in blue, and Milo in green) in identifying B, NK, T, or momac cells in (A) Patient 695 and (B) Patient 706. Since the difference was calculated in reference to the performance of existing methods, positive values represent superiority of scDeepJointClust over existing methods. FPR difference of scDeepJointCluster vs. existing methods (Leiden in red, Louvain in blue, and Milo in green) in identifying B, NK, T, or momac cells in (C) Patient 695 and (D) Patient 706. Since the difference was calculated in reference to the performance of existing methods, negative values represent superiority of scDeepJointClust over existing methods.

4.5 Conclusions

In this project, we introduced scDeepJointClust, a novel approach for clustering cells into subtypes or states by refining an initial clustering result with the consideration of metadata indicating the biological conditions. We further conducted multiple experiments using simulation data generated in diverse scenarios and biological data of different contexts (melanoma and NSCLC tumors treated with immunotherapy treatments) to demonstrate superiority of scDeepJointClust over existing methods, raising three important implications as follows. First, scDeepJointClust considers previously unexplored information during the clustering process such as whether cells were derived from tumor or normal samples, representing an innovative approach to refine the performance of existing sophisticated methods. Secondly, scDeepJointClust harnesses the power of state-of-the-art gene-expression-based clustering methods, incorporating their sophistication and accuracy. This ensures that scDeepJointClust stays at the cutting edge of performance by leveraging the advancements in gene-expression-based clustering techniques. Third, by employing the DNN method to embed and train on both types of information, gene expression and the biological condition, scDeepJointClust successfully captures and models the nonlinear relationship in how cell states are defined with such data.

However, scDeepJointClust also calls for further investigation to tackle some methodological and analytical limitations. A methodological limitation is that our approach capitalizes on enriched biological conditions in the cells of the same type, which may not be advantageous for identifying cell types that do not exhibit such enrichments. However, considering that the primary focus of immunologic studies lies in identifying differentially abundant cell types between conditions, we believe that scDeepJointClust effectively addresses this specific interest. From an analytical standpoint, it is important to acknowledge that the evaluation of true positive

rate (TPR) and false positive rate (FPR) against the original cluster IDs in the NSCLC data might not be entirely precise. Although the original IDs are expected to closely approximate the true cell types, as they rely not only on gene expression profiles but also on cell epitope information, the extent to which cell identity can be accurately learned from single-cell RNA-Seq or CITE-Seq data remains an ongoing area of study. Therefore, our evaluation based on the original cell type IDs may not precisely reflect the true cell identification performance. Nonetheless, when combined with our simulation results, which showcases superiority of scDeepJointClust over the simulated truth, the results together strongly suggest that it outperforms other existing methods in accurately defining true cell clusters.

In summary, we developed scDeepJointClust, which identifies cellular states that are differentially abundant between biological conditions. Identifying those cellular states is utmost important because it provides crucial insights into the functional and molecular diversity within a tissue or organism. However, the current formulation of this problem employs a two-step approach, which can potentially lead to suboptimal solutions. In contrast, scDeepJointClust tackles this problem by transforming it into a joint-learning problem and leveraging a DNN-based approach. This innovative methodology facilitates more accurate identification of cellular states, thus providing valuable insights into the underlying functional and molecular diversity associated with important pathobiology. By harnessing these advantages, the application of scDeepJointClust holds significant promise for advancing our understanding of cellular states and their implications in complex biological systems.

Bibliography

[1] G. Sliwoski, S. Kothiwale, J. Meiler, and J. Edward W. Lowe, "Computational Methods in Drug Discovery," *Pharmacol. Rev.*, vol. 66, no. 1, pp. 334 LP – 395, Jan. 2014, doi: 10.1124/pr.112.007336.

[2] A. V Sadybekov and V. Katritch, "Computational approaches streamlining drug discovery," *Nature*, vol. 616, no. 7958, pp. 673–685, 2023, doi: 10.1038/s41586-023-05905-z.

[3] F. Shao and Z. Shen, "How can artificial neural networks approximate the brain?," *Front. Psychol.*, vol. 13, 2023, doi: 10.3389/fpsyg.2022.970214.

[4] X. Lin, X. Li, and X. Lin, "A Review on Applications of Computational Methods in Drug Screening and Design," *Molecules*, vol. 25, no. 6. 2020, doi: 10.3390/molecules25061375.

[5] P. Bogdan, "Taming the Unknown Unknowns in Complex Systems: Challenges and Opportunities for Modeling, Analysis and Control of Complex (Biological) Collectives," *Front. Physiol.*, vol. 10, 2019, doi: 10.3389/fphys.2019.01452.

[6] A. Aderem, "Systems Biology: Its Practice and Challenges," *Cell*, vol. 121, no. 4, pp. 511–513, May 2005, doi: 10.1016/j.cell.2005.04.020.

[7] C. E. Hmelo-Silver and R. Azevedo, "Understanding Complex Systems: Some Core Challenges," *J. Learn. Sci.*, vol. 15, no. 1, pp. 53–61, Jun. 2006, [Online]. Available: http://www.jstor.org/stable/25473509.

[8] C. Xu and S. A. Jackson, "Machine learning and complex biological data," *Genome Biol.*, vol. 20, no. 1, p. 76, 2019, doi: 10.1186/s13059-019-1689-0.

[9] Paul Williams, "Challenges and Opportunities in Computational Biology and Systems Biology," *Int. J. Swarm Intell. Evol. Comput.*, vol. 12, no. 2, 2023, doi: 10.35248/2090-4908.23.12.304.

[10] H. C. Yeo and K. Selvarajoo, "Machine learning alternative to systems biology should not solely depend on data," *Brief. Bioinform.*, vol. 23, no. 6, p. bbac436, Nov. 2022, doi: 10.1093/bib/bbac436.

[11] M. Koutrouli, E. Karatzas, D. Paez-Espino, and G. A. Pavlopoulos, "A Guide to Conquer the Biological Network Era Using Graph Theory," *Front. Bioeng. Biotechnol.*, vol. 8, 2020, doi: 10.3389/fbioe.2020.00034.

[12] X. Zhu, M. Gerstein, and M. Snyder, "Getting connected: analysis and principles of biological networks," *Genes Dev.*, vol. 21, no. 9, pp. 1010–1024, 2007, doi: 10.1101/gad.1528707.

[13] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943.

[14] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, “Deep learning for computational biology.,” *Mol. Syst. Biol.*, vol. 12, no. 7, p. 878, Jul. 2016, doi: 10.15252/msb.20156651.

[15] T. Ching *et al.*, “Opportunities and obstacles for deep learning in biology and medicine.,” *J. R. Soc. Interface*, vol. 15, no. 141, Apr. 2018, doi: 10.1098/rsif.2017.0387.

[16] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning,” *Nat. Biotechnol.*, vol. 33, no. 8, pp. 831–838, 2015, doi: 10.1038/nbt.3300.

[17] S. Min, B. Lee, and S. Yoon, “Deep learning in bioinformatics,” *Brief. Bioinform.*, vol. 18, no. 5, pp. 851–869, 2016, doi: 10.1093/bib/bbw068.

[18] R. Poplin *et al.*, “A universal SNP and small-indel variant caller using deep neural networks,” *Nat. Biotechnol.*, vol. 36, no. 10, pp. 983–987, 2018, doi: 10.1038/nbt.4235.

[19] M. L. Kuijjer, J. N. Paulson, P. Salzman, W. Ding, and J. Quackenbush, “Cancer subtype identification using somatic mutation data,” *Br. J. Cancer*, vol. 118, no. 11, pp. 1492–1501, 2018, doi: 10.1038/s41416-018-0109-7.

[20] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” *Nat. Rev. Genet.*, vol. 16, no. 6, pp. 321–332, 2015, doi: 10.1038/nrg3920.

[21] J. P. Higgins, “Nonlinear systems in medicine.,” *Yale J. Biol. Med.*, vol. 75, pp. 247–260, 2002.

[22] C. Trefois, P. M. A. Antony, J. Goncalves, A. Skupin, and R. Balling, “Critical transitions in chronic disease: transferring concepts from ecology to systems medicine,” *Curr. Opin. Biotechnol.*, vol. 34, pp. 48–55, 2015, doi: https://doi.org/10.1016/j.copbio.2014.11.020.

[23] N. Naik *et al.*, “Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains,” *Nat. Commun.*, vol. 11, no. 1, p. 5727, 2020, doi: 10.1038/s41467-020-19334-3.

[24] G. Lebedeva, A. Yamaguchi, S. P. Langdon, K. Macleod, and D. J. Harrison, “A model of estrogen-related gene expression reveals non-linear effects in transcriptional response to tamoxifen,” *BMC Syst. Biol.*, vol. 6, no. 1, p. 138, 2012, doi: 10.1186/1752-0509-6-138.

[25] M. Perera and C. Tsokos, “A Statistical Model with Non-Linear Effects and Non-Proportional Hazards for Breast Cancer Survival Analysis,” *Adv. Breast Cancer Res.*, vol. 07, pp. 65–89, Jan. 2018, doi: 10.4236/abcr.2018.71005.

[26] A. J. Brooks, J. W. Wooh, K. A. Tunny, and M. J. Waters, “Growth hormone receptor; mechanism of action,” *Int. J. Biochem. Cell Biol.*, vol. 40, no. 10, pp. 1984–1989, 2008, doi: https://doi.org/10.1016/j.biocel.2007.07.008.

[27] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search, 2nd Edition*, vol. 81. 2000.

[28] X. Zhang *et al.*, "Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information," *Bioinformatics*, vol. 28, no. 1, pp. 98–104, Jan. 2012, doi: 10.1093/bioinformatics/btr626.

[29] M. H. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann, "Predicting causal effects in large-scale systems from observational data," *Nat. Methods*, vol. 7, no. 4, pp. 247–248, 2010, doi: 10.1038/nmeth0410-247.

[30] T. D. Le *et al.*, "Inferring microRNA–mRNA causal regulatory relationships from expression data," *Bioinformatics*, vol. 29, no. 6, pp. 765–771, Mar. 2013, doi: 10.1093/bioinformatics/btt048.

[31] J. Zhang *et al.*, "Inferring condition-specific miRNA activity from matched miRNA and mRNA expression data," *Bioinformatics*, vol. 30, no. 21, pp. 3070–3077, Nov. 2014, doi: 10.1093/bioinformatics/btu489.

[32] J. Zhang *et al.*, "Identifying direct miRNA–mRNA causal regulatory relationships in heterogeneous data," *J. Biomed. Inform.*, vol. 52, pp. 438–447, 2014, doi: https://doi.org/10.1016/j.jbi.2014.08.005.

[33] C. Silverstein, S. Brin, R. Motwani, and J. Ullman, "Scalable Techniques for Mining Causal Structures," *Data Min. Knowl. Discov.*, vol. 4, no. 2, pp. 163–192, 2000, doi: 10.1023/A:1009891813863.

[34] B. Andrews, J. Ramsey, and G. F. Cooper, "Learning High-dimensional Directed Acyclic Graphs with Mixed Data-types," in *Proceedings of Machine Learning Research*, 2019, vol. 104, pp. 4–21, [Online]. Available: http://proceedings.mlr.press/v104/andrews19a.html.

[35] G. Schwarz, "Estimating the Dimension of a Model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, May 1978, [Online]. Available: http://www.jstor.org/stable/2958889.

[36] D. M. Chickering, "Optimal Structure Identification With Greedy Search.," *J. Mach. Learn. Res.*, vol. 3, no. null, pp. 507–554, Jan. 2003, doi: 10.1162/153244303321897717.

[37] Y. Yu, J. Chen, T. Gao, and M. Yu, "DAG-GNN: DAG structure learning with graph neural networks," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 12395–12406, 2019.

[38] X. Zheng, C. Dan, and B. Aragam, "Learning Sparse Nonparametric DAGs," vol. 108, 2020.

[39] M. E. H. Hammond *et al.*, "American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version).," *Arch. Pathol. Lab. Med.*, vol. 134, no. 7, pp. e48-72, Jul. 2010, doi: 10.1043/1543-2165-134.7.e48.

[40] M. Kormaksson, L. J. Kelly, X. Zhu, S. Haemmerle, L. Pricop, and D. Ohlssen, “Sequential knockoffs for continuous and categorical predictors: With application to a large psoriatic arthritis clinical trial pool,” *Stat. Med.*, vol. 40, no. 14, pp. 3313–3328, 2021.

[41] C. Mayr and D. P. Bartel, “Widespread shortening of 3’UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.,” *Cell*, vol. 138, no. 4, pp. 673–84, Aug. 2009, doi: 10.1016/j.cell.2009.06.016.

[42] M. Chen *et al.*, “3 ′ UTR lengthening as a novel mechanism in regulating cellular senescence,” *Genome Res.*, vol. 28, no. 3, pp. 285–294, 2018, doi: 10.1101/gr.224451.117.Freely.

[43] G. P. Dimri *et al.*, “A biomarker that identifies senescent human cells in culture and in aging skin in vivo.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 20, pp. 9363–7, 1995, doi: DOI 10.1073/pnas.92.20.9363.

[44] R. A. Busuttil, M. Rubio, M. E. T. Dollé, J. Campisi, and J. Vijg, “Oxygen accelerates the accumulation of mutations during the senescence and immortalization of murine cells in culture.,” *Aging Cell*, vol. 2, no. 6, pp. 287–294, 2003, doi: 10.1046/j.1474-9728.2003.00066.x.

[45] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, “The hallmarks of aging,” *Cell*, vol. 153, no. 6, 2013, doi: 10.1016/j.cell.2013.05.039.

[46] D. Muñoz-Espín and M. Serrano, “Cellular senescence: From physiology to pathology,” *Nat. Rev. Mol. Cell Biol.*, vol. 15, no. 7, pp. 482–496, 2014, doi: 10.1038/nrm3823.

[47] Z. Xia *et al.*, “Dynamic Analyses of Alternative Polyadenylation from RNA- Seq Reveal Landscape of 3 ’ UTR Usage Across 7 Tumor Types,” *Nat. Commun.*, pp. 1–38, 2014.

[48] Y. Xiang *et al.*, “Comprehensive Characterization of Alternative Polyadenylation in Human Cancer,” vol. 110, no. November 2017, pp. 1–11, 2018, doi: 10.1093/jnci/djx223.

[49] L. Salmena, L. Poliseno, Y. Tay, L. Kats, and P. P. Pandolfi, “A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?,” *Cell*, vol. 146, no. 3, pp. 353–8, Aug. 2011, doi: 10.1016/j.cell.2011.07.014.

[50] P. Sumazin *et al.*, “An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma.,” *Cell*, vol. 147, no. 2, pp. 370–81, Oct. 2011, doi: 10.1016/j.cell.2011.09.041.

[51] H. J. Park, S. Kim, and W. Li, “Model-based analysis of competing- endogenous pathways (MACPath) in human cancers,” *PLoS Comput. Biol.*, vol. 22, no. 14, 2018, doi: 10.1371/journal.pcbi.1006074.

[52] H. J. Park *et al.*, “3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk,” *Nat. Genet.*, vol. 50, pp. 783–789, 2018, doi: 10.1038/s41588-018-0118-8.

[53] S. Kim, Y. Bai, Z. Fan, B. Diergaarde, G. C. Tseng, and H. J. Park, "Alternative Polyadenylation Regulates Patient-specific Tumor Growth by Individualizing the MicroRNA Target Site Landscape," *bioRxiv*, p. 601518, Jan. 2019, doi: 10.1101/601518.

[54] S. Tsutsui, S. Ohno, S. Murakami, Y. Hachitanda, and S. Oda, "Prognostic value of epidermal growth factor receptor (EGFR) and its relationship to the estrogen receptor status in 1029 patients with breast cancer," *Breast Cancer Res. Treat.*, vol. 71, no. 1, pp. 67–75, 2002.

[55] M. Sheikh, G. M, P. P, F. JA, and R. H, "Why are estrogen-receptor-negative breast cancers more aggressive t," *Invasion Metastasis*, vol. 14, no. 1–6, pp. 329–36, 1994.

[56] A. Pergamenschikov *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2002, doi: 10.1038/35021093.

[57] M. Goldman *et al.*, "The UCSC Cancer Genomics Browser: update 2013.," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D949-54, Jan. 2013, doi: 10.1093/nar/gks1008.

[58] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.," *Cell*, vol. 120, no. 1, pp. 15–20, Jan. 2005, doi: 10.1016/j.cell.2004.12.035.

[59] G. L. Papadopoulos, M. Reczko, V. a Simossis, P. Sethupathy, and A. G. Hatzigeorgiou, "The database of experimentally supported targets: a functional update of TarBase.," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D155-8, Jan. 2009, doi: 10.1093/nar/gkn809.

[60] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, "miRecords: An integrated resource for microRNA-target interactions," *Nucleic Acids Res.*, vol. 37, no. November 2008, pp. 105–110, 2009, doi: 10.1093/nar/gkn851.

[61] S.-D. Hsu *et al.*, "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions.," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D78-85, Jan. 2014, doi: 10.1093/nar/gkt1266.

[62] H. Dvinge *et al.*, "The shaping and functional consequences of the microRNA landscape in breast cancer.," *Nature*, vol. 497, no. 7449, pp. 378–82, May 2013, doi: 10.1038/nature12108.

[63] M. P. Hamilton *et al.*, "Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif.," *Nat. Commun.*, vol. 4, p. 2730, Jan. 2013, doi: 10.1038/ncomms3730.

[64] J. Y. Lee, I. Yeh, J. Y. Park, and B. Tian, "PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes," *Nucleic Acids Res.*, vol. 35, no. suppl_1, pp. D165–D168, Jan. 2007, doi: 10.1093/nar/gkl870.

[65] T. L. Bailey *et al.*, "MEME Suite: Tools for motif discovery and searching," *Nucleic Acids*

Res., vol. 37, no. SUPPL. 2, pp. 202–208, 2009, doi: 10.1093/nar/gkp335.

[66] E. Eisenberg and E. Levanon, "Human housekeeping genes are compact," *TRENDS Genet.*, vol. 19, no. 7, pp. 362–365, 2003, doi: 10.1016/S0168-9525(03)00140-9.

[67] E. Eisenberg and E. Y. Levanon, "Human housekeeping genes, revisited.," *Trends Genet.*, vol. 29, no. 10, pp. 569–74, Oct. 2013, doi: 10.1016/j.tig.2013.05.010.

[68] K. Chawla, S. Tripathi, L. Thommesen, A. Lægreid, and M. Kuiper, "TFcheckpoint: A curated compendium of specific DNA-binding RNA polymerase II transcription factors," *Bioinformatics*, vol. 29, no. 19, pp. 2519–2520, 2013, doi: 10.1093/bioinformatics/btt432.

[69] T. Davoli *et al.*, "Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome.," *Cell*, vol. 155, no. 4, pp. 948–962, Oct. 2013, doi: 10.1016/j.cell.2013.10.011.

[70] U. Ala *et al.*, "Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 18, pp. 7154–9, Apr. 2013, doi: 10.1073/pnas.1222509110.

[71] R. Gera *et al.*, "Identifying network structure similarity using spectral graph theory," *Appl. Netw. Sci.*, vol. 3, no. 1, p. 2, 2018, doi: 10.1007/s41109-017-0042-3.

[72] C. C. Friedel and R. Zimmer, "Inferring topology from clustering coefficients in protein-protein interaction networks," *BMC Bioinformatics*, vol. 15, pp. 1–15, 2006, doi: 10.1186/1471-2105-7-519.

[73] E. Dalgıç, Ö. Konu, Z. S. Öz, and C. Chan, "Lower connectivity of tumor coexpression networks is not specific to cancer," *In Silico Biol.*, vol. 13, no. 1–2, pp. 41–53, 2019, doi: 10.3233/ISB-190472.

[74] H. Chen *et al.*, "A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples," *Cell*, vol. 173, no. 2, pp. 386-399.e12, Apr. 2018, doi: 10.1016/j.cell.2018.03.027.

[75] G. Altay, M. Asim, F. Markowetz, and D. E. Neal, "Differential C3NET reveals disease networks of direct physical interactions," *BMC Bioinformatics*, vol. 12, no. 1, p. 296, 2011, doi: 10.1186/1471-2105-12-296.

[76] P. Curmi *et al.*, "Overexpression of stathmin in breast carcinomas points out to highly proliferative tumours," *Br. J. Cancer*, vol. 82, no. 1, pp. 142–150, 2000.

[77] S. Obayashi *et al.*, "Stathmin1 expression is associated with aggressive phenotypes and cancer stem cell marker expression in breast cancer patients," *Int. J. Oncol.*, vol. 51, no. 3, pp. 781–790, 2017, doi: 10.3892/ijo.2017.4085.

[78] K. Krishnan *et al.*, "miR-139-5p is a regulator of metastatic pathways in breast cancer.," *RNA*, vol. 19, no. 12, pp. 1767–80, Dec. 2013, doi: 10.1261/rna.042143.113.

[79] D. Vaught, D. M. Brantley-Sieders, and J. Chen, “Eph receptors in breast cancer: Roles in tumor promotion and tumor suppression,” *Breast Cancer Research*. 2008, doi: 10.1186/bcr2207.

[80] Y. Tay, J. Rinn, and P. P. Pandolfi, “The multilayered complexity of ceRNA crosstalk and competition,” *Nature*, vol. 505, no. 7483, pp. 344–352, Jan. 2014, doi: 10.1038/nature12986.

[81] E. P. Bastos *et al.*, “MicroRNAs Discriminate Familial from Sporadic Non-BRCA1/2 Breast Carcinoma Arising in Patients ≤35 Years,” *PLoS One*, vol. 9, no. 7, p. e101656, Jul. 2014, doi: 10.1371/journal.pone.0101656.

[82] C. P. Masamha *et al.*, “CFIm25 links alternative polyadenylation to glioblastoma tumour suppression.,” *Nature*, vol. 510, no. 7505, pp. 412–416, May 2014, doi: 10.1038/nature13261.

[83] A. Curinha, S. Oliveira Braz, I. Pereira-Castro, A. Cruz, and A. Moreira, “Implications of polyadenylation in health and disease,” *Nucleus*, vol. 5, no. 6, pp. 508–519, 2014, doi: 10.4161/nucl.36360.

[84] G. Hong *et al.*, “Genes Dysregulated to Different Extent or Oppositely in Estrogen Receptor-Positive and Estrogen Receptor-Negative Breast Cancers,” *PLoS One*, vol. 8, no. 7, p. e70017, 2013, doi: 10.1371/journal.pone.0070017.

[85] M. C. Alles *et al.*, “Meta-Analysis and Gene Set Enrichment Relative to ER Status Reveal Elevated Activity of MYC and E2F in the ‘Basal’ Breast Cancer Subgroup,” *PLoS One*, vol. 4, no. 3, p. e4710, 2009, doi: 10.1371/journal.pone.0004710.

[86] M. C. Abba *et al.*, “Gene expression signature of estrogen receptor α status in breast cancer,” *BMC Genomics*, vol. 6, pp. 1–13, 2005, doi: 10.1186/1471-2164-6-37.

[87] D. K. Biswas, a. P. Cruz, E. Gansberger, and a. B. Pardee, “Epidermal growth factor-induced nuclear factor kappa B activation: A major pathway of cell-cycle progression in estrogen-receptor negative breast cancer cells,” *Proc. Natl. Acad. Sci.*, vol. 97, no. 15, pp. 8542–8547, Jul. 2000, doi: 10.1073/pnas.97.15.8542.

[88] D. Fuckar *et al.*, “VEGF expression is associated with negative estrogen receptor status in patients with breast cancer,” *Int. J. Surg. Pathol.*, vol. 14, no. 1, pp. 49–55, 2006, doi: 10.1177/106689690601400109.

[89] S. Javanmoghadam, Z. Weihua, K. K. Hunt, and K. Keyomarsi, “Estrogen receptor alpha is cell cycle-regulated and regulates the cell cycle in a ligand-dependent fashion,” *Cell Cycle*, vol. 15, no. 12, pp. 1579–1590, 2016, doi: 10.1080/15384101.2016.1166327.

[90] S. Paruthiyil, H. Parmar, V. Kerekatte, G. R. Cunha, G. L. Firestone, and D. C. Leitman, “Estrogen Receptor ␣ Inhibits Human Breast Cancer Cell Proliferation and Tumor Formation by Causing a G 2 Cell Cycle Arrest,” pp. 423–428, 2004.

[91] D. C. Henley, J. S. Foster, J. Wimalasena, P. Seth, and A. Bukovsky, “Multifaceted Regulation of Cell Cycle Progression by Estrogen: Regulation of Cdk Inhibitors and Cdc25A Independent of Cyclin D1-Cdk4 Function,” *Mol. Cell. Biol.*, vol. 21, no. 3, pp. 794–810, 2002, doi: 10.1128/mcb.21.3.794-810.2001.

[92] Y. Tay *et al.*, “Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs.,” *Cell*, vol. 147, no. 2, pp. 344–57, Oct. 2011, doi: 10.1016/j.cell.2011.09.029.

[93] T. Tuersong, L. Li, Z. Abulaiti, and S. Feng, “Comprehensive analysis of the aberrantly expressed lncRNA-associated ceRNA network in breast cancer,” *Mol. Med. Rep.*, vol. 19, no. 6, pp. 4697–4710, Jun. 2019, doi: 10.3892/mmr.2019.10165.

[94] Y. Wei, Z. Chang, C. Wu, Y. Zhu, K. Li, and Y. Xu, “Identification of potential cancer-related pseudogenes in lung adenocarcinoma based on ceRNA hypothesis,” *Oncotarget*, vol. 8, no. 35, pp. 59036–59047, Aug. 2017, doi: 10.18632/oncotarget.19933.

[95] W. Bouhaddioui, P. R. Provost, and Y. Tremblay, “Identification of Most Stable Endogenous Control Genes for MicroRNA Quantification in the Developing Mouse Lung,” *PLoS One*, vol. 9, no. 11, p. e111855, Nov. 2014.

[96] D. Cheng, Y. Xiang, L. Ji, and X. Lu, “Competing endogenous RNA interplay in cancer : mechanism , methodology , and perspectives,” 2015, doi: 10.1007/s13277-015-3093-z.

[97] The Gene Ontology Consortium, “The Gene Ontology Resource: 20 years and still GOing strong,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D330–D338, Nov. 2018, doi: 10.1093/nar/gky1055.

[98] S. Kim, H. J. Park, X. Cui, and D. Zhi, “Collective effects of long-range DNA methylations predict gene expressions and estimate phenotypes in cancer,” *Sci. Rep.*, vol. 10, no. 1, p. 3920, 2020, doi: 10.1038/s41598-020-60845-2.

[99] S. Kim, Y. Bai, Z. Fan, B. Diergaarde, G. C. Tseng, and H. J. Park, “The microRNA target site landscape is a novel molecular feature associating alternative polyadenylation with immune evasion activity in breast cancer,” *Brief. Bioinform.*, vol. 00, no. July, pp. 1–10, 2020, doi: 10.1093/bib/bbaa191.

[100] Z. Fan, S. Kim, Y. Bai, B. Diergaarde, and H. J. Park, “3′-UTR Shortening Contributes to Subtype-Specific Cancer Growth by Breaking Stable ceRNA Crosstalk of Housekeeping Genes,” *Frontiers in Bioengineering and Biotechnology*, vol. 8. p. 334, 2020.

[101] A. J. Sedgewick, J. D. Ramsey, P. Spirtes, C. Glymour, and P. V. Benos, “Mixed Graphical Models for Causal Analysis of Multi-modal Variables,” *CoRR*, vol. abs/1704.0. 2017.

[102] P.-L. Loh and P. Bühlmann, “High-Dimensional Learning of Linear Causal Networks via Inverse Covariance Estimation,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3065–3105, Jan. 2014.

[103] J. D. Lee and T. J. Hastie, “Structure learning of mixed graphical models,” *Journal of Machine Learning Research*, vol. 31. PMLR, pp. 388–396, 2013.

[104] R. Cui, P. Groot, and T. Heskes, *Copula PC Algorithm for Causal Discovery from Mixed Data*, vol. 9852. 2016.

[105] A. J. Sedgewick, I. Shi, R. M. Donovan, and P. V. Benos, “Learning mixed graphical models with separate sparsity parameters and stability-based model selection,” *BMC Bioinformatics*, vol. 17, no. 5, p. S175, 2016, doi: 10.1186/s12859-016-1039-0.

[106] R. F. Barber and E. J. Candes, “Controlling the false discovery rate via knockoffs,” *Ann. Stat.*, vol. 43, no. 5, pp. 2055–2085, 2015, doi: 10.1214/15-AOS1337.

[107] Z. Fan *et al.*, “Deep neural networks with knockoff features identify nonlinear causal relations and estimate effect sizes in complex biological systems,” *Gigascience*, vol. 12, p. giad044, Jan. 2023, doi: 10.1093/gigascience/giad044.

[108] S. G. Bottcher and C. Dethlefsen, “Learning Bayesian Networks with Mixed Variables,” *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, vol. R3. PMLR, pp. 13–20, 2011.

[109] V. Romero, R. Rumí, and A. Salmerón, “Learning hybrid Bayesian networks using mixtures of truncated exponentials,” *Int. J. Approx. Reason.*, vol. 42, no. 1, pp. 54–68, 2006, doi: https://doi.org/10.1016/j.ijar.2005.10.004.

[110] H. E. Kyburg, “Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference by Judea Pearl,” *Journal of Philosophy*, vol. 88, no. 8. Morgan kaufmann, pp. 434–437, 1991, doi: 10.5840/jphil199188844.

[111] M. Koivisto and K. Sood, “Exact Bayesian Structure Discovery in Bayesian Networks,” *J. Mach. Learn. Res.*, vol. 5, pp. 549–573, Dec. 2004.

[112] T. Silander and P. Myllymäki, “A Simple Approach for Finding the Globally Optimal Bayesian Network Structure,” *ArXiv*, vol. abs/1206.6, Jun. 2006.

[113] T. Jaakkola, D. Sontag, A. Globerson, and M. M. B. T.-P. of the T. I. C. on A. I. and Statistics, “Learning Bayesian Network Structure using LP Relaxations,” vol. 9. PMLR, pp. 358–365.

[114] J. Cussens, “Bayesian Network Learning with Cutting Planes,” in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 2011, pp. 153–160.

[115] C. Yuan, B. Malone, and X. Wu, “Learning optimal Bayesian networks using A*search,” *IJCAI International Joint Conference on Artificial Intelligence*. pp. 2186–2191, Jul. 16, 2011, doi: 10.5591/978-1-57735-516-8/IJCAI11-364.

[116] T. Gao and D. Wei, “Parallel {B}ayesian Network Structure Learning,” in *Proceedings of the 35th International Conference on Machine Learning*, 2018, vol. 80, pp. 1685–1694,

[Online]. Available: https://proceedings.mlr.press/v80/gao18b.html.

[117] E. C. Neto, M. P. Keller, A. D. Attie, and B. S. Yandell, “CAUSAL GRAPHICAL MODELS IN SYSTEMS GENETICS: A UNIFIED FRAMEWORK FOR JOINT INFERENCE OF CAUSAL NETWORK AND GENETIC ARCHITECTURE FOR CORRELATED PHENOTYPES.,” *Ann. Appl. Stat.*, vol. 4, no. 1, pp. 320–339, Mar. 2010, doi: 10.1214/09-aoas288.

[118] W. Kruijer *et al.*, “Reconstruction of Networks with Direct and Indirect Genetic Effects.,” *Genetics*, vol. 214, no. 4, pp. 781–807, Apr. 2020, doi: 10.1534/genetics.119.302949.

[119] A. Yazdani, A. Yazdani, A. Samiei, and E. Boerwinkle, “Generating a robust statistical causal structure over 13 cardiovascular disease risk factors using genomics data.,” *J. Biomed. Inform.*, vol. 60, pp. 114–119, Apr. 2016, doi: 10.1016/j.jbi.2016.01.012.

[120] A. Yazdani, A. Yazdani, A. Saniei, and E. Boerwinkle, “A causal network analysis in an observational study identifies metabolomics pathways influencing plasma triglyceride levels.,” *Metabolomics*, vol. 12, no. 6, p. 104, 2016, doi: 10.1007/s11306-016-1045-2.

[121] A. Yazdani, A. Yazdani, T. A. Bowman, F. Marotta, J. P. Cooke, and A. Samiei, “Arachidonic acid as a target for treating hypertriglyceridemia reproduced by a causal network analysis and an intervention study.,” *Metabolomics : Official journal of the Metabolomic Society*, vol. 14, no. 6. United States, p. 78, May 2018, doi: 10.1007/s11306-018-1368-2.

[122] A. Yazdani *et al.*, “Genome analysis and pleiotropy assessment using causal networks with loss of function mutation and metabolomics.,” *BMC Genomics*, vol. 20, no. 1, p. 395, May 2019, doi: 10.1186/s12864-019-5772-4.

[123] S. Triantafillou, V. Lagani, C. Heinze-Deml, A. Schmidt, J. Tegner, and I. Tsamardinos, “Predicting Causal Relationships from Biological Data: Applying Automated Causal Discovery on Mass Cytometry Data of Human Immune Cells.,” *Sci. Rep.*, vol. 7, no. 1, p. 12724, Oct. 2017, doi: 10.1038/s41598-017-08582-x.

[124] D. Rothenhäusler, C. Heinze, J. Peters, and N. Meinshausen, “BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions,” *Adv. Neural Inf. Process. Syst.*, vol. 2015-Janua, no. 1, pp. 1513–1521, 2015.

[125] A. J. Sedgewick *et al.*, “Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis.,” *Bioinformatics*, vol. 35, no. 7, pp. 1204–1212, Apr. 2019, doi: 10.1093/bioinformatics/bty769.

[126] S. Nie, D. D. Maua, C. P. de Campos, and Q. Ji, “Advances in Learning Bayesian Networks of Bounded Treewidth,” in *Advances in Neural Information Processing Systems*, 2014, vol. 27.

[127] M. Scanagatta, C. P. de Campos, G. Corani, and M. Zaffalon, “Learning Bayesian Networks with Thousands of Variables,” in *Advances in Neural Information Processing Systems*,

2015, vol. 28.

[128] E. Y.-J. Chen, Y. Shen, A. Choi, and A. Darwiche, "Learning Bayesian networks with ancestral constraints," in *Advances in Neural Information Processing Systems*, 2016, vol. 29.

[129] K. Rantanen, A. Hyttinen, and M. Järvisalo, "Discovering causal graphs with cycles and latent confounders: An exact branch-and-bound approach," *Int. J. Approx. Reason.*, vol. 117, pp. 29–49, 2020, doi: https://doi.org/10.1016/j.ijar.2019.10.009.

[130] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous Optimization for Structure Learning," in *Advances in Neural Information Processing Systems*, 2018, vol. 31.

[131] Y. Y. Lu, Y. Fan, J. Lv, and W. S. Noble, "Deeppink: Reproducible feature selection in deep neural networks," *Advances in Neural Information Processing Systems*, vol. 2018-Decem. pp. 8676–8686, Sep. 04, 2018.

[132] G. D. Wu *et al.*, "Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes," *Science (80-.).*, vol. 334, no. 6052, pp. 105–108, May 2011, [Online]. Available: http://www.jstor.org/stable/23059312.

[133] D. J. Slamon *et al.*, "Studies of the HER-2/neu Proto-Oncogene in Human Breast and Ovarian Cancer," *Science (80-.).*, vol. 244, no. 4905, pp. 707–712, May 1989, [Online]. Available: http://www.jstor.org/stable/1703358.

[134] J. Chen and H. Li, "Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis," *Ann. Appl. Stat.*, vol. 7, no. 1, pp. 418–442, Mar. 2013, doi: 10.1214/12-AOAS592.

[135] W. Lin, P. Shi, R. Feng, and H. Li, "Variable selection in regression with compositional covariates," *Biometrika*, vol. 101, no. 4, pp. 785–797, Dec. 2014, doi: 10.1093/biomet/asu031.

[136] J. A. Carcillo *et al.*, "A Multicenter Network Assessment of Three Inflammation Phenotypes in Pediatric Sepsis-Induced Multiple Organ Failure.," *Pediatr. Crit. care Med. a J. Soc. Crit. Care Med. World Fed. Pediatr. Intensive Crit. Care Soc.*, vol. 20, no. 12, pp. 1137–1146, Dec. 2019, doi: 10.1097/PCC.0000000000002105.

[137] B. Goldstein, B. Giroir, and A. Randolph, "International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics.," *Pediatr. Crit. care Med. a J. Soc. Crit. Care Med. World Fed. Pediatr. Intensive Crit. Care Soc.*, vol. 6, no. 1, pp. 2–8, Jan. 2005, doi: 10.1097/01.PCC.0000149131.72248.E6.

[138] A. Villeneuve, J.-S. Joyal, F. Proulx, T. Ducruet, N. Poitras, and J. Lacroix, "Multiple organ dysfunction syndrome in critically ill children: clinical value of two lists of diagnostic criteria.," *Ann. Intensive Care*, vol. 6, no. 1, p. 40, Dec. 2016, doi: 10.1186/s13613-016-0144-6.

[139] A. Yazdani, A. Yazdani, A. Samiei, and E. Boerwinkle, “Identification, analysis, and interpretation of a human serum metabolomics causal network in an observational study.,” *J. Biomed. Inform.*, vol. 63, pp. 337–343, Oct. 2016, doi: 10.1016/j.jbi.2016.08.017.

[140] A. F. Agarap, “Deep Learning using Rectified Linear Units (ReLU).” arXiv, 2018, doi: 10.48550/ARXIV.1803.08375.

[141] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, vol. 9, pp. 249–256, [Online]. Available: https://proceedings.mlr.press/v9/glorot10a.html.

[142] C. Glymour, K. Zhang, and P. Spirtes, “Review of Causal Discovery Methods Based on Graphical Models ,” *Frontiers in Genetics* , vol. 10. p. 524, 2019.

[143] E. Candes, Y. Fan, L. Janson, and J. Lv, “Panning for Gold: Model-free Knockoffs for High-dimensional Controlled Variable Selection,” *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 80, Oct. 2016, doi: 10.1111/rssb.12265.

[144] E. Candès, Y. Fan, L. Janson, and J. Lv, “Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection,” *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 80, no. 3, pp. 551–577, Jun. 2018, doi: https://doi.org/10.1111/rssb.12265.

[145] W. Hardle and T. M. Stoker, “Investigating Smooth Multiple Regression by the Method of Average Derivatives,” *J. Am. Stat. Assoc.*, vol. 84, no. 408, pp. 986–995, Mar. 1989, doi: 10.2307/2290074.

[146] H. Ichimura, “Semiparametric least squares (SLS) and weighted SLS estimation of single-index models,” *J. Econom.*, vol. 58, no. 1, pp. 71–120, 1993, doi: https://doi.org/10.1016/0304-4076(93)90114-K.

[147] R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand, “Generalized Partially Linear Single-Index Models,” *J. Am. Stat. Assoc.*, vol. 92, no. 438, pp. 477–489, Mar. 1997, doi: 10.2307/2965697.

[148] L. Wang and L. Yang, “SPLINE ESTIMATION OF SINGLE-INDEX MODELS,” *Stat. Sin.*, vol. 19, no. 2, pp. 765–783, Mar. 2009, [Online]. Available: http://www.jstor.org/stable/24308855.

[149] Y. Qin *et al.*, “Four computable 24-hour pediatric sepsis phenotypes have different inflammation profiles and heterogeneous outcome with anti-inflammatory therapies,” *Crit. Care*, 2022.

[150] C. B. Crayne, S. Albeituni, K. E. Nichols, and R. Q. Cron, “The Immunology of Macrophage Activation Syndrome.,” *Front. Immunol.*, vol. 10, p. 119, 2019, doi: 10.3389/fimmu.2019.00119.

[151] I. Ushach and A. Zlotnik, “Biological role of granulocyte macrophage colony-stimulating

factor (GM-CSF) and macrophage colony-stimulating factor (M-CSF) on cells of the myeloid lineage," *J. Leukoc. Biol.*, vol. 100, no. 3, pp. 481–489, Sep. 2016, doi: 10.1189/jlb.3RU0316-144R.

[152] S. L. Deshmane, S. Kremlev, S. Amini, and B. E. Sawaya, "Monocyte Chemoattractant Protein-1 (MCP-1): An Overview," *J. Interf. Cytokine Res.*, vol. 29, no. 6, pp. 313–326, May 2009, doi: 10.1089/jir.2008.0027.

[153] L. Zhu, Q. Zhao, T. Yang, W. Ding, and Y. Zhao, "Cellular metabolism and macrophage functional polarization.," *Int. Rev. Immunol.*, vol. 34, no. 1, pp. 82–100, Jan. 2015, doi: 10.3109/08830185.2014.969421.

[154] A. Dige *et al.*, "Soluble CD163, a Specific Macrophage Activation Marker, is Decreased by Anti-TNF-α Antibody Treatment in Active Inflammatory Bowel Disease," *Scand. J. Immunol.*, vol. 80, no. 6, pp. 417–423, Dec. 2014, doi: https://doi.org/10.1111/sji.12222.

[155] N. Rittig, M. Svart, N. Jessen, N. Møller, H. J. Møller, and H. Grønbæk, "Macrophage activation marker sCD163 correlates with accelerated lipolysis following LPS exposure: a human-randomised clinical trial," *Endocr. Connect.*, vol. 7, no. 1, pp. 107–114, 2018, doi: 10.1530/EC-17-0296.

[156] A. V Finn *et al.*, "Hemoglobin directs macrophage differentiation and prevents foam cell formation in human atherosclerotic plaques," *J. Am. Coll. Cardiol.*, vol. 59, no. 2, pp. 166–177, Jan. 2012, doi: 10.1016/j.jacc.2011.10.852.

[157] S. Fleming *et al.*, "Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies," *Lancet*, vol. 377, no. 9770, pp. 1011–1018, 2011, doi: https://doi.org/10.1016/S0140-6736(10)62226-X.

[158] D. R. Jury, "Serum creatinine concentration in children: normal values for sex and age," *N. Z. Med. J.*, vol. 90, no. 649, pp. 453–456, 1979, [Online]. Available: http://europepmc.org/abstract/MED/294520.

[159] W. T. Shearer *et al.*, "Lymphocyte subsets in healthy children from birth through 18 years of age: The pediatric AIDS clinical trials group P1009 study," *J. Allergy Clin. Immunol.*, vol. 112, no. 5, pp. 973–980, 2003, doi: https://doi.org/10.1016/j.jaci.2003.07.003.

[160] M. W. Merx and C. Weber, "Sepsis and the Heart," *Circulation*, vol. 116, no. 7, pp. 793–802, Aug. 2007, doi: 10.1161/CIRCULATIONAHA.106.678359.

[161] L. Ma *et al.*, "Role of interleukin-6 to differentiate sepsis from non-infectious systemic inflammatory response syndrome," *Cytokine*, vol. 88, pp. 126–135, 2016, doi: https://doi.org/10.1016/j.cyto.2016.08.033.

[162] C. Mitaka, "Clinical laboratory differentiation of infectious versus non-infectious systemic inflammatory response syndrome," *Clin. Chim. Acta*, vol. 351, no. 1, pp. 17–29, 2005, doi: https://doi.org/10.1016/j.cccn.2004.08.018.

[163] K. Nakanishi, “Unique Action of Interleukin-18 on T Cells and Other Immune Cells,” *Frontiers in Immunology*, vol. 9. p. 763, 2018.

[164] J. R. Schoenborn and C. B. Wilson, “Regulation of interferon-gamma during innate and adaptive immune responses.,” *Adv. Immunol.*, vol. 96, pp. 41–101, 2007, doi: 10.1016/S0065-2776(07)96002-2.

[165] A. C. Stanley and P. Lacy, “Pathways for Cytokine Secretion,” *Physiology*, vol. 25, no. 4, pp. 218–229, Aug. 2010, doi: 10.1152/physiol.00017.2010.

[166] W. J. Leonard and J.-X. Lin, “Cytokine receptor signaling pathways,” *J. Allergy Clin. Immunol.*, vol. 105, no. 5, pp. 877–888, 2000, doi: https://doi.org/10.1067/mai.2000.106899.

[167] W. Tate *et al.*, “Molecular Mechanisms of Neuroinflammation in ME/CFS and Long COVID to Sustain Disease and Promote Relapses ,” *Frontiers in Neurology* , vol. 13. 2022, [Online]. Available: https://www.frontiersin.org/article/10.3389/fneur.2022.877772.

[168] L. Zhao *et al.*, “Sepsis-Associated Encephalopathy: Insight into Injury and Pathogenesis,” *CNS & Neurological Disorders - Drug Targets*, vol. 20, no. 2. pp. 112–124, 2021, doi: http://dx.doi.org/10.2174/1871527319999201117122158.

[169] G. F. Weber, S. Schlautkötter, S. Kaiser-Moore, F. Altmayr, B. Holzmann, and H. Weighardt, “Inhibition of interleukin-22 attenuates bacterial load and organ failure during acute polymicrobial sepsis.,” *Infect. Immun.*, vol. 75, no. 4, pp. 1690–1697, Apr. 2007, doi: 10.1128/IAI.01564-06.

[170] S. Manicka, K. Johnson, M. Levin, and D. Murrugarra, “Biological regulatory networks are less nonlinear than expected by chance,” *bioRxiv*, p. 2021.12.22.473903, Jan. 2021, doi: 10.1101/2021.12.22.473903.

[171] T. Kapitaniak and S. Jafari, “Nonlinear effects in life sciences,” *Eur. Phys. J. Spec. Top.*, vol. 227, no. 7, pp. 693–696, 2018, doi: 10.1140/epjst/e2018-800104-6.

[172] R. Stoof and Á. Goñi-Moreno, “Modelling co-translational dimerization for programmable nonlinearity in synthetic biology,” *J. R. Soc. Interface*, vol. 17, no. 172, p. 20200561, Nov. 2020, doi: 10.1098/rsif.2020.0561.

[173] H. Blankson, J. A. Stakkestad, H. Fagertun, E. Thom, J. Wadstein, and O. Gudmundsen, “Conjugated linoleic acid reduces body fat mass in overweight and obese humans.,” *J. Nutr.*, vol. 130, no. 12, pp. 2943–2948, Dec. 2000, doi: 10.1093/jn/130.12.2943.

[174] C.-M. Chiu *et al.*, “Systematic analysis of the association between gut flora and obesity through high-throughput sequencing and bioinformatics approaches.,” *Biomed Res. Int.*, vol. 2014, p. 906168, 2014, doi: 10.1155/2014/906168.

[175] M. Vanhala *et al.*, “Serum omega-6 polyunsaturated fatty acids and the metabolic syndrome: a longitudinal population-based cohort study.,” *Am. J. Epidemiol.*, vol. 176, no.

3, pp. 253–260, Aug. 2012, doi: 10.1093/aje/kwr504.

[176] L. Pimpin, S. Jebb, L. Johnson, J. Wardle, and G. L. Ambrosini, “Dietary protein intake is associated with body mass index and weight up to 5 y of age in a prospective cohort of twins.,” *Am. J. Clin. Nutr.*, vol. 103, no. 2, pp. 389–397, Feb. 2016, doi: 10.3945/ajcn.115.118612.

[177] S. Rabot *et al.*, “High fat diet drives obesity regardless the composition of gut microbiota in mice,” *Sci. Rep.*, vol. 6, no. 1, p. 32484, 2016, doi: 10.1038/srep32484.

[178] Q. Yang, “Gain weight by ‘going diet?’ Artificial sweeteners and the neurobiology of sugar cravings: Neuroscience 2010,” *Yale J. Biol. Med.*, vol. 83, no. 2, pp. 101–108, Jun. 2010, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20589192.

[179] Y. Yun *et al.*, “Comparative analysis of gut microbiota associated with body mass index in a large Korean cohort.,” *BMC Microbiol.*, vol. 17, no. 1, p. 151, Jul. 2017, doi: 10.1186/s12866-017-1052-0.

[180] D. N. Reeds, B. S. Mohammed, S. Klein, C. B. Boswell, and V. L. Young, “Metabolic and structural effects of phosphatidylcholine and deoxycholate injections on subcutaneous fat: a randomized, controlled trial.,” *Aesthetic Surg. J.*, vol. 33, no. 3, pp. 400–408, Mar. 2013, doi: 10.1177/1090820X13478630.

[181] Y.-S. Kuang *et al.*, “Connections between the human gut microbiome and gestational diabetes mellitus.,” *Gigascience*, vol. 6, no. 8, pp. 1–12, Aug. 2017, doi: 10.1093/gigascience/gix058.

[182] Y. J. Yang, Y. J. Kim, Y. K. Yang, J. Y. Kim, and O. Kwon, “Dietary flavan-3-ols intake and metabolic syndrome risk in Korean adults.,” *Nutr. Res. Pract.*, vol. 6, no. 1, pp. 68–77, Feb. 2012, doi: 10.4162/nrp.2012.6.1.68.

[183] D. C. Koboldt *et al.*, “Comprehensive molecular portraits of human breast tumours,” *Nature*, vol. 490, no. 7418, pp. 61–70, 2012, doi: 10.1038/nature11412.

[184] T. Pascual *et al.*, “A Pathology-Based Combined Model to Identify PAM50 Non-luminal Intrinsic Disease in Hormone Receptor-Positive HER2-Negative Breast Cancer ,” *Frontiers in Oncology* , vol. 9. 2019, [Online]. Available: https://www.frontiersin.org/article/10.3389/fonc.2019.00303.

[185] T. O. Nielsen *et al.*, “A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor–Positive Breast Cancer,” *Clin. Cancer Res.*, vol. 16, no. 21, pp. 5222–5232, Oct. 2010, doi: 10.1158/1078-0432.CCR-10-1282.

[186] M. Rossing *et al.*, “Clinical implications of intrinsic molecular subtypes of breast cancer for sentinel node status,” *Sci. Rep.*, vol. 11, no. 1, p. 2259, 2021, doi: 10.1038/s41598-021-81538-4.

[187] E. A. Mittendorf, J. M. S. Bartlett, D. L. Lichtensztajn, and S. Chandarlapaty, "Incorporating Biology Into Breast Cancer Staging: American Joint Committee on Cancer, Eighth Edition, Revisions and Beyond," *Am. Soc. Clin. Oncol. Educ. B.*, no. 38, pp. 38–46, May 2018, doi: 10.1200/EDBK_200981.

[188] A. A. Onitilo, J. M. Engel, R. T. Greenlee, and B. N. Mukesh, "Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival," *Clin. Med. Res.*, vol. 7, no. 1–2, pp. 4–13, Jun. 2009, doi: 10.3121/cmr.2009.825.

[189] X. Dai, A. Chen, and Z. Bai, "Integrative investigation on breast cancer in ER, PR and HER2-defined subgroups using mRNA and miRNA expression profiling," *Sci. Rep.*, vol. 4, no. 1, p. 6566, 2014, doi: 10.1038/srep06566.

[190] K. P. Harden and K. L. Klump, "Introduction to the Special Issue on Gene-Hormone Interplay," *Behav. Genet.*, vol. 45, no. 3, pp. 263–267, 2015, doi: 10.1007/s10519-015-9717-7.

[191] M. W. Coolen *et al.*, "Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity.," *Nat. Cell Biol.*, vol. 12, no. 3, pp. 235–246, Mar. 2010, doi: 10.1038/ncb2023.

[192] S. Ashida *et al.*, "Integrated analysis reveals critical genomic regions in prostate tumor microenvironment associated with clinicopathologic phenotypes.," *Clin. cancer Res. an Off. J. Am. Assoc. Cancer Res.*, vol. 18, no. 6, pp. 1578–1587, Mar. 2012, doi: 10.1158/1078-0432.CCR-11-2535.

[193] P. Flaherty, P. Wiratchotisatian, J. A. Lee, Z. Tang, and A. C. Trapp, "MAP Clustering under the Gaussian Mixture Model via Mixed Integer Nonlinear Optimization." arXiv, 2019, doi: 10.48550/ARXIV.1911.04285.

[194] S. Kim *et al.*, "Expression Quantitative Trait Methylation Analysis Reveals Methylomic Associations With Gene Expression in Childhood Asthma," *Chest*, vol. 158, no. 5, pp. 1841–1856, Nov. 2020, doi: 10.1016/j.chest.2020.05.601.

[195] R. Karki *et al.*, "Synergism of TNF-α and IFN-γ Triggers Inflammatory Cell Death, Tissue Damage, and Mortality in SARS-CoV-2 Infection and Cytokine Shock Syndromes," *Cell*, vol. 184, no. 1, pp. 149-168.e17, 2021, doi: https://doi.org/10.1016/j.cell.2020.11.025.

[196] R. N. Gomes *et al.*, "Bacterial clearance in septic mice is modulated by MCP-1/CCL2 and nitric oxide.," *Shock*, vol. 39, no. 1, pp. 63–69, Jan. 2013, doi: 10.1097/SHK.0b013e31827802b5.

[197] A.-C. Villani *et al.*, "Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors," *Science (80-.).*, vol. 356, no. 6335, p. eaah4573, 2017, doi: 10.1126/science.aah4573.

[198] M. J. Heinrich *et al.*, "Endogenous double-stranded Alu RNA elements stimulate IFN-responses in relapsing remitting multiple sclerosis.," *J. Autoimmun.*, vol. 100, pp. 40–51,

Jun. 2019, doi: 10.1016/j.jaut.2019.02.003.

[199] B. Mahata *et al.*, "Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis.," *Cell Rep.*, vol. 7, no. 4, pp. 1130–1142, May 2014, doi: 10.1016/j.celrep.2014.04.011.

[200] L. Alberti-Servera *et al.*, "Single-cell RNA sequencing reveals developmental heterogeneity among early lymphoid progenitors," *EMBO J.*, vol. 36, no. 24, pp. 3619–3633, 2017, doi: https://doi.org/10.15252/embj.201797105.

[201] T. Kreisel *et al.*, "Dynamic microglial alterations underlie stress-induced depressive-like behavior and suppressed neurogenesis.," *Mol. Psychiatry*, vol. 19, no. 6, pp. 699–709, Jun. 2014, doi: 10.1038/mp.2013.155.

[202] S. Davidson *et al.*, "Single-Cell RNA Sequencing Reveals a Dynamic Stromal Niche That Supports Tumor Growth.," *Cell Rep.*, vol. 31, no. 7, p. 107628, May 2020, doi: 10.1016/j.celrep.2020.107628.

[203] A. M. Leader *et al.*, "Single-cell analysis of human non-small cell lung cancer lesions refines tumor classification and patient stratification," *Cancer Cell*, vol. 39, no. 12, pp. 1594-1609.e12, 2021, doi: https://doi.org/10.1016/j.ccell.2021.10.009.

[204] L. Li, Z. Li, Y. Liu, and Q. Hong, "Deep joint learning for language recognition," *Neural Networks*, vol. 141, pp. 72–86, 2021, doi: https://doi.org/10.1016/j.neunet.2021.03.026.

[205] Y. Li, X. Tian, X. Shen, and D. Tao, "Classification and Representation Joint Learning via Deep Networks," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, {IJCAI-17}*, 2017, pp. 2215–2221, doi: 10.24963/ijcai.2017/308.

[206] P. Gundogdu, C. Loucera, I. Alamo-Alvarez, J. Dopazo, and I. Nepomuceno, "Integrating pathway knowledge with deep neural networks to reduce the dimensionality in single-cell RNA-seq data," *BioData Min.*, vol. 15, no. 1, p. 1, 2022, doi: 10.1186/s13040-021-00285-4.

[207] S. Srinivasan, A. Leshchyk, N. T. Johnson, and D. Korkin, "A hybrid deep clustering approach for robust cell type profiling using single-cell RNA-seq data.," *RNA*, vol. 26, no. 10, pp. 1303–1319, Oct. 2020, doi: 10.1261/rna.074427.119.

[208] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982, doi: 10.1109/TIT.1982.1056489.

[209] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011, [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html.

[210] R. Cannoodt, W. Saelens, L. Deconinck, and Y. Saeys, "Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells," *Nat. Commun.*, vol. 12, no. 1, p. 3942, 2021.

[211] E. Dann, N. C. Henderson, S. A. Teichmann, M. D. Morgan, and J. C. Marioni, “Differential abundance testing on single-cell data using k-nearest neighbor graphs,” *Nat. Biotechnol.*, vol. 40, no. 2, pp. 245–253, 2022, doi: 10.1038/s41587-021-01033-z.

[212] Y. Hao *et al.*, “Integrated analysis of multimodal single-cell data,” *Cell*, vol. 184, no. 13, pp. 3573-3587.e29, 2021, doi: https://doi.org/10.1016/j.cell.2021.04.048.

[213] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[214] F. A. Wolf, P. Angerer, and F. J. Theis, “SCANPY: large-scale single-cell gene expression data analysis,” *Genome Biol.*, vol. 19, no. 1, p. 15, 2018, doi: 10.1186/s13059-017-1382-0.

[215] V. A. Traag, L. Waltman, and N. J. van Eck, “From Louvain to Leiden: guaranteeing well-connected communities,” *Sci. Rep.*, vol. 9, no. 1, Mar. 2019, doi: 10.1038/s41598-019-41695-z.

[216] D. Aran *et al.*, “Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage,” *Nat. Immunol.*, vol. 20, no. 2, pp. 163–172, 2019, doi: 10.1038/s41590-018-0276-y.

[217] Joost H.A. Martens and Hendrik G. Stunnenberg, “BLUEPRINT: mapping human blood cell epigenomes,” *Haematologica*, vol. 98, no. 10 SE-Editorials, pp. 1487–1489, Oct. 2013, doi: 10.3324/haematol.2013.094243.

[218] I. Tirosh *et al.*, “Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma.,” *Nature*, vol. 539, no. 7628, pp. 309–313, Nov. 2016, doi: 10.1038/nature20123.

[219] P. Brennecke *et al.*, “Accounting for technical noise in single-cell RNA-seq experiments,” *Nat. Methods*, vol. 10, no. 11, pp. 1093–1095, 2013, doi: 10.1038/nmeth.2645.

[220] T. Stuart *et al.*, “Comprehensive Integration of Single-Cell Data.,” *Cell*, vol. 177, no. 7, pp. 1888-1902.e21, Jun. 2019, doi: 10.1016/j.cell.2019.05.031.

[221] V. Y. Kiselev *et al.*, “SC3: consensus clustering of single-cell RNA-seq data,” *Nat. Methods*, vol. 14, no. 5, pp. 483–486, 2017, doi: 10.1038/nmeth.4236.

[222] R. Tibshirani, T. Hastie, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction : with 200 Full-color Illustrations*. Springer, 2001.

[223] A. Gelman and E. Loken, “The Statistical Crisis in Science,” *Am. Sci.*, vol. 102, p. 460, Nov. 2014, doi: 10.1511/2014.111.460.

[224] J. P. A. Ioannidis, “Correction: Why Most Published Research Findings Are False,” *PLOS Med.*, vol. 19, no. 8, p. 1, 2022, doi: 10.1371/journal.pmed.1004085.

[225] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[226] M. Sade-Feldman *et al.*, "Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma," *Cell*, vol. 175, no. 4, pp. 998-1013.e20, Nov. 2018, doi: 10.1016/j.cell.2018.10.038.

[227] J. C. Martin *et al.*, "Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy.," *Cell*, vol. 178, no. 6, pp. 1493-1508.e20, Sep. 2019, doi: 10.1016/j.cell.2019.08.008.

[228] D. Lambrechts *et al.*, "Phenotype molding of stromal cells in the lung tumor microenvironment.," *Nat. Med.*, vol. 24, no. 8, pp. 1277–1289, Aug. 2018, doi: 10.1038/s41591-018-0096-5.