UNCOVERING HIDDEN CONNECTIONS AND FUNCTIONAL MODULES VIA pyPARAGON: A HYBRID APPROACH FOR NETWORK CONTEXTUALIZATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

MÜSLÜM KAAN ARICI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF MEDICAL INFORMATICS

JANUARY 2024

Approval of the thesis:

UNCOVERING HIDDEN CONNECTIONS AND FUNCTIONAL MODULES VIA pyPARAGON: A HYBRID APPROACH FOR NETWORK CONTEXTUALIZATION

Submitted by MÜSLÜM KAAN ARICI in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Health Informatics Department, Middle East Technical University** by,

Date: *22.01.2024*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : MÜSLÜM KAAN ARICI

Signature : _____

ABSTRACT

UNCOVERING HIDDEN CONNECTIONS AND FUNCTIONAL MODULES VIA pyPARAGON: A HYBRID APPROACH FOR NETWORK CONTEXTUALIZATION

Arici, Müslüm Kaan

Ph.D., Department of Department of Health Informatics

Supervisor: Asst. Prof. Dr. Aybar Can Acar

Co-Supervisor: Assoc. Prof. Dr. Nurcan Tunçbağ

January 2024, 115 pages

State-of-the-art omics technologies provide molecular insights into various biological contexts, such as disease states, patients, and drug perturbations. Network inference and reconstruction methods utilize several omics datasets to create context-based networks that reveal the interactions of biomolecules and the functioning of cells. We compared the coverage of reference networks in several categories of prior knowledge, such as pathways, three-dimensional structures of interactions, and publication counts of genes/proteins to detect constraints in reference networks. Additionally, we examined the limitations of reconstruction algorithms by inferring signaling pathways. Contextualized network inference has several challenging issues: i) Hits from omics datasets are sparse in reference networks. ii) Interpretation methods can miss hidden knowledge that connects significant hits in omics datasets while evaluating multi-omics datasets. iii) Well-studied proteins in reference networks come along with bias in contextualization. iv) Highly connected nodes, or hubs, cause unspecific and noisy interactions in inferred networks. To overcome these challenges, we developed pyPARAGON (PAgeRAnk-flux on Graphlet-guided network for multi-Omics data integratioN). Combining network propagation with graphlets, pyPARAGON also improves precision and reduces the presence of non-specific interactions in contextualized networks. We tested the performance of pyPARAGON by reconstructing cancer-associated signaling pathways and setting contextual models of different cancer types. Moreover, pyPARAGON has promising performance in case studies such as tumor-specific networks with significant biological processes and contextualized neurodevelopmental disorders and cancer models, including signal strength on their shared pathways.

Keywords: Network reconstruction, graphlets, data integration, interactome, disease modeling

ÖZ

GİZLİ ETKİLEŞİMLER VE FONKSİYONEL MODÜLLERİN HİBRİT BİR AĞ BAĞLAMSALLAŞTIRMA ARACI pyPARAGON İLE AÇIĞA ÇIKARILMASI

Arıcı, Müslüm Kaan

Doktora, Sağlık Bilişimi Bölümü

Tez Yöneticisi: Dr. Öğr. Üyesi Aybar Can Acar

Eş-Danışman: Doç. Dr. Nurcan Tunçbağ

OCAK 2024, 115 sayfa

En son omiks teknolojileri, hastalık durumları, hastalar ve ilaç bozuklukları gibi çeşitli biyolojik bağlamlarda moleküler bilgi sağlamaktadır. Ağ varsayım ve yeniden yapılandırma yöntemleri, belirli biyomoleküler etkileşimler ve hücresel süreçlerle bağlamlı ağlar oluşturmak için birkaç omiks veri kümesi kullanır. Referans ağlarının kapsamını, referans ağlarında kısıtlamaları tespit etmek için önceki bilginin yolları, üç boyutlu etkileşim yapısı ve gen/protein yayın sayımları gibi çeşitli kategorilerde karşılaştırdık. Yeniden yapılandırma algoritmalarının sınırlamalarını sinyal yolaklarını inceledik. Bağlamsallaştırılmış ağ oluşturmak birkaç zorlu meseleye sahiptir: (i) Omiks çıktılar referans ağlara göre çok küçük kalırlar. (ii) Değerlendirme yöntemleri, multi-omiks verisetlerini değerlendirirken omiks veri kümelerinde önemli çıktıları bağlayan gizli bilgiyi kaçırabilirler. iii) Referans ağlardaki çok çalışılan proteinler bağlamsallaştırmada yanlılığı beraberinde getirir. (iv) Yüksek sayıda bağlantıya sahip düğümler, ya da hublar, oluşturulan ağlarda özel olmayan veya yanlış etkileşimlerin olmasına yol açar. Bu zorluklarla başa çıkmak için pyPARAGON (PAgeRAnk-flux on Graphlet-guided network for multi-Omics data integratioN)'u geliştirdik. PyPARAGON, ağ yayılmasını graphlets ile birleştirerek aynı zamanda hassasiyetini artırırken ve bağlam dışı ağlarda spesifik olmayan etkileşimlerin varlığını azaltmaktadır. pyPARAGON'un performansını, kanserle ilişkili sinyal yollarını yeniden yapılandırarak ve farklı kanser türlerinin bağlamsal modelleri ile test ettik. Ayrıca, pyPARAGON, önemli biyolojik süreçlerle tümör spesifik ağları ve ortak yollarındaki sinyal gücü de dahil olmak üzere bağlamlı nörolojik gelişim bozuklukları ve kanser modelleri gibi durum çalışmalarında umut verici bir performans sergiledi.

Anahtar Sözcükler: Ağların yeniden inşası, gaflet, veri entegrasyonu, interaktom, hastalık modelleme

To My Running Mates

ACKNOWLEDGMENTS

Firstly, I would like to sincerely thank my supervisors, Assoc. Prof. Nurcan Tunçbağ and Asst. Prof. Aybar Can ACAR, whose encouragement, companionship, and steadfast support throughout the process facilitated the successful completion of my dissertation.

I am grateful to Prof. Ruth Nussinov for her mentorship, for imparting her extensive expertise, and for constantly offering assistance. I would like to share my appreciation with the members of Ruth Nussinov's research group for their fruitful scientific discussions. In particular, I would like to acknowledge the valuable contributions of Dr. Chung-Jung Tsai and Dr. Hyunbum Jang in the case studies. I express my gratitude to my thesis committee members, Assoc. Prof. Elif Sürer, Assoc. Prof. Mehmet Tan, Asst. Prof. Burçak Otlu-Sarıtaş and Assoc. Prof. Ceren Sucular for their meticulous examination and valuable feedback.

I thank Dr. Bengi R. Yavuz and H. Cansu Demirel for their valuable contributions to the case studies. I am also thankful to the Network Modeling Research Group members and everyone I had the opportunity to discuss with this dissertation.

Furthermore, I am truly indebted to my family for their academic and psychological support. I am grateful to my parents, Satı and Kamil Arıcı, and my sisters, Ilknur Erarslan and Aynur Boytwhite, for their endless support in every aspect of life. This thesis would not have been possible without their ongoing support. I would like to extend my appreciation to our running crew, who have become like a second family to me, namely Ahmet Kavalcı, Asst Prof. Serap Emil, Dr. Sadun Tanışer, Demet Göl, Betül Ağaç Obuz, and Erhan Obuz for their precious support and encouragement.

Projects and research in this thesis have been funded in whole or in part with the TUBITAK fund under project number 121C292 and with the national support program BIDEB 2211A.

TABLE OF CONTENTS

ABSTRACT iv
ÖZ v
DEDICATION vi
ACKNOWLEDGMENTS vii
TABLE OF CONTENTS viii
LIST OF TABLES xi
LIST OF FIGURES xii
LIST OF ABBREVIATIONS xiii
CHAPTERS
1. INTRODUCTION 1
1.1. Background 1
1.2. Motivation 3
1.3. Contributions of the study 4
1.4. Organization of the dissertation 6
2. LITERATURE REVIEW 9
2.1. Biological networks 9
2.2. Contextualization of biological networks 11
2.2.1. Data integration 12
2.2.2. Contextualization approaches 16
2.2.3. Interpretations of context-specific networks 21
2.3. Graphlets 25
2.3.1. Graphlet-based metrics 25
2.3.2. Graphlet motifs 28
3. PERFORMANCE ASSESSMENT OF THE NETWORK RECONSTRUCTION APPROACHES ON VARIOUS INTERACTOMES 29
3.1. Methods 29
3.1.1. Reference interactomes 29
3.1.2. Interactome comparison metrics 30

3.1.3. Network reconstruction methods 32
3.1.4. Performance analysis 35
3.2. Results 37
3.2.1. Systematic evaluation of reference human interactomes 37
3.2.2. Performance of network reconstruction algorithms 43
3.2.3. Reconstruction of the notch pathway 47
4. pyPARAGON: COMBINING NETWORK PROPAGATION WITH GRAPHLETS TO INTEGRATE MULTI-OMICS DATA 49
4.1. Methods 50
4.1.1. Overview of pyPARAGON as a hybrid network inference framework 50
4.1.2. Network inference tools 51
4.1.3. Interactomes and datasets 53
4.1.4. Performance assessment of network inference tools 53
4.2. Results 55
4.2.1. Network trimming via graphlets improves the reference networks 55
4.2.2. Performance of pyPARAGON on the reconstruction of cancer signaling pathways 59
4.2.3. Network-based modeling of cancer types 61
4.2.4. Running time analysis of pyPARAGON 63
5. IMPLEMENTATION OF pyPARAGON 65
5.1. Methods: 66
5.1.1. Case Study 1: Contextualization of breast cancer samples 66
5.1.2. Case study 2: Contextualization of neurodevelopmental disorders and cancers 68
5.2. Results: 71
5.2.1. Case Study 1: Tumor-specific network inference reveals hidden commonalities across tumors 71
5.2.2. Case Study 2: Disease-specific networks identifies shared pathways . 75
6. DISCUSSION 81
6.1. Evaluation of the network reconstruction approaches on various interactomes 81
6.2. pyPARAGON unveiled hidden knowledge by contextualizing networks 83

REFERENCES 87
APPENDICES 109
APPENDIX A 109
APPENDIX B 111
CURRICULUM VITAE 115

LIST OF TABLES

Table 1: Statistics of interactomes 30
Table 2: Tunning ranges of parameter sets in PageRank flux (PRF), heat diffusion flux (HDF), and prize-collecting Steiner Forest. 37
Table 3: Topological features of reference networks..... 57
Table 4: Expression profiles of diseases 70

LIST OF FIGURES

Figure 1: The conceptual overview of multi-omics data integration approaches and their applications 13
Figure 2: Integrative network-based approaches 15
Figure 3: The conceptual representation of reconstruction algorithms 19
Figure 4: 14 Automorphism orbits for nine graphlets 26
Figure 5: An example of a 5-node dummy graphlet. 27
Figure 6: A comparative analysis of the reference interactomes 39
Figure 7: Correlation between publication counts and degrees in the context of interactomes 42
Figure 8: Coverages of known structurally known and curated interactions 43
Figure 9: Principal Component Analysis (PCA) over edge-based and node-based scores 45
Figure 10: Performance assessment of each interactome and method in pathway reconstruction 46
Figure 11: Reconstructed Notch pathway 48
Figure 12: The overview of pyPARAGON A) pyPARAGON 51
Figure 13: Graphlet-guided networks (GGN) optimize reference networks 58
Figure 14: Graphlet-guided network trims reference interactomes 59
Figure 15: Outperformance of pyPARAGON over Omics Integrator2 and PathLinker 61
Figure 16: Performance of contextualized cancer-specific networks 63
Figure 17: Running time graph based on the initial node size 64
Figure 18: Running time graph based on the network size 64
Figure 19: Contextualization and downstream analysis of tumor-specific networks 67
Figure 20: A conceptual representation of network comparison analysis between NDDs and cancer. 69
Figure 21: Example of active modules in a tumor-specific network constructed 71
Figure 22: Stratification of tumors and associated biological processes with patient clusters 73
Figure 23: Survival analysis and cluster-specific KEGG pathways. 74
Figure 24: Drug-module interaction network of a patient 75
Figure 25: ASD- and breast cancer-specific networks regulating common pathways. 78
Figure 26: Differential TFs drive to proliferation in cancer and differentiation in ASD 80

LIST OF ABBREVIATIONS

APSP	The All-Pairs Shortest Path
ASD	Autism Spectrum Disorder
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast Invasive Carcinoma
CDGs	Cancer Driver Genes
CDKs	Cyclin-Dependent Kinases
CGI	Cancer Genome Interpreter
CPTAC	The Clinical Proteomic Tumor Analysis Consortium
DepMap	Cancer Dependency Map
EA	Enrichment Analysis
ES	The Expression Score
ESCA	Esophageal Carcinoma
FPR	False Positive Rate
GDD	The Graphlet Degree Distribution
GDV	The Graphlet Degree Vector
GMM	The Gaussian Mixture Model
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HD	Heat Diffusion
HDF	Heat Diffusion with Flux
HNSC	Head and Neck Squamous Cell Carcinoma
ICGC	The International Cancer Genome Consortium
LPP	The Locality Preserving Projections
LUSC	Lung Squamous Cell Carcinoma
MCC	Matthew's Correlation Coefficient
MI	MINT-Inspired
NDD	Neurodevelopmental Disorders
ORA	Overrepresentation Analysis
PAAD	Pancreatic Adenocarcinoma
PCSF	Prize-Collecting Steiner Forest

PPIs	Protein-Protein Interactions
PPR	Personalized PageRank
PRAD	Prostate Adenocarcinoma
PRF	Personalized Pagerank with Flux
PS	The Propensity Score
pyPARAGON	Pagerank-Flux on Graphlet-Guided Network for Multi-Omics Data Integration
scATAC-seq	Single-Cell Assays for Transposase-Accessible Chromatin Using Sequencing
scRNA-seq	Single-Cell RNA-Sequencing
STFs	Specific Transcription Factors
t-SNE	T-Distributed Stochastic Neighbour Embedding
TCGA	The Cancer Genome Atlas
TFs	Transcription Factors
UCEC	Uterine Corpus Endometrial Carcinoma

CHAPTER 1

INTRODUCTION

1.1. Background

The central dogma of molecular biology governs biological processes such as the cell cycle, DNA, replication, chromosome packing, epigenetic alterations, transcription, posttranscriptional alterations, translation, and posttranslational modifications[1,2]. Signal transductions, including transcription factors, protein complexes, and enzymes, tightly regulate the transmission of genetic information from DNA to the phenotype. In order to understand how biological processes work in cell machines, high-throughput methods are generative to measure changes at different molecular levels of the central dogma, such as gene, transcription, protein, and metabolite levels[3,4].

High-throughput methods identify all detectable biomolecules at their molecular level and produce omics datasets named by their molecular classes, such as genomics for the genome, epigenomics for the epigenome, transcriptomics for the transcriptome, proteomics for the proteome, and metabolomics for the metabolome. With a high number of entries, omics datasets need computational methods to clarify unanswered problems by propagating experimental information. However, a single level of data could not efficiently explain biological issues such as biological processes, disease onsets, and drug perturbations. Rather than just looking at one type of omics data, integrative approaches to multi-omics datasets help us understand how different molecules interact and how biological systems work. Multi-omics data integration methods are a holistic approach that covers the central dogma and systematically uncovers the relationships between omics entities[5,6].

Integration methods also recruit prior information from various databases, such as reference interactomes, annotated biological processes, cellular pathways, and disease associations[4,7]. Integrated databases provide explanatory annotations by checking their

reliability against various data sources. Several sources of prior information include protein-protein interactions (PPIs)[8,9], genetic variants[10,11], and protein/gene expression profiles[12,13]. For example, pathways comprise cataloged and sequential PPIs explaining signal transduction or other cellular processes. The regulatomes, composed of regulator elements and their interactions, and the interactomes, which consist of PPIs, hold the pathway information through molecular interactions[14,15]. Looking at prior knowledge in omics datasets, overrepresentation and enrichment analysis, causal inference, and network reconstruction can give a deep understanding of biological problems[7,16–18].

Recent studies have utilized omics datasets to identify drug-response patterns and perturbation signatures in diseases, predict biomarkers, and develop therapeutics or patient stratification. For example, the integration of multi-omics data provides biological insights for drug repurposing[19] transcriptional dysregulation of pathways in Alzheimer's disease[20], comprehensive molecular profiles of SARS-CoV-2 infection to propose drug candidates[21], pathway modulation by drugs in breast cancer cell lines, and novel alcoholism-related genes[22]. Furthermore, several efforts have been made to investigate complex diseases in the same set of tumors, patients, and perturbations. Because of their multifactorial nature, complex diseases are mainly modeled with context-specific solutions using omics datasets. The development of high-throughput methods has come with the accumulation of data on various disease types, mainly cancer. Just some examples of contextual databases are the Cancer Genome Atlas (TCGA)[23], the International Cancer Genome Consortium (ICGC)[24], Clinical Proteomic Tumor Analysis Consortium (CPTAC)[25], TARGET[26] for pediatric malignancies, Cancer Cell Line Encyclopedia (CCLE)[27] and denovo-db for germline *de novo* variants[28].

The integrative methods include machine learning strategies, network-based applications, statistical methods, or a mix of multiple approaches. Ultimately, multi-omics datasets are transformed into interpretable knowledge, including significant variants, pathways, biological processes, drug targets, and stratified patients. However, large-scale datasets at different molecular levels require a cost-efficient solution. Thus, these methods specifically focus on dimensionality reduction and data exploration by

recruiting omics data sequentially and simultaneously[29,30]. Formerly, omics datasets were separately optimized and improved by the previous omics dataset based on the information flow in the central dogma[31–33]. However, this process may bring about a loss of sensitivity for weak signals. In the latter, integration methods optimize all omics datasets simultaneously, causing overfitting problems and information loss.

Network-based algorithms, such as Steiner tree/forest[34], random walk[35,36], or heat diffusion[17,37], transform the list of genes/proteins from omics datasets into context-specific networks using the topological properties of the reference network. These approaches cover disease-associated subnetworks, signature modules, or biomarkers. The coverage of reference interactomes is another challenging issue due to false positives and negatives in network inference[38]. Thus, interactions in these databases have been scored with various calculations, considering experimental detection methods, the number of publications, interologs, and many other gold-standard properties[8,39,40]. However, the well-studied genes/proteins in cancer research cause a bias in the interactomes[41]. Therefore, some hub proteins—proteins that have a high number of interactions—have highly scored interactions. Several network-based algorithms penalize these hub proteins and identify their context-specific interactions to overcome the biased interactions in the interactomes[34,42,43]. Beyond network inference, meaningful communities in contextualized networks illuminate the detailed functionalities of genes/proteins and predict the rewired regions of biological processes. The state of art solutions in network-based algorithms may provide not only a precise context-specific network but also interpretable outputs.

1.2. Motivation

Omics technologies are the most advanced ways to find dysregulated and changed signaling parts in various biological contexts, such as disease onsets, proceedings, patients, and drug classes. Network-based methods can propagate signals from omics datasets, infer context-specific networks, and identify their functional communities when integrating multi-omics datasets. These methods mainly utilize the global and local properties of networks and annotations in reference networks. Recruited algorithms, pipelines, and prior knowledge affect their results and interpretations. Their performances alter on different datasets due to a wide variety of limitations.

Thus, our ultimate aim was to release a novel, open-source tool that processed omics hits and delivered interpretable molecular associations by working in balance under the limitations.

The first motivation for this Ph.D. study was to apprehend the constraints of input datasets, covering omics datasets and reference networks. A straightforward analysis of a list of genes from omics or previous studies may predict all direct neighbors as associated genes. However, such a naive analysis could predict noisy genes and miss associated genes due to incomplete data in the reference networks. Omics datasets reflect temporal hits in samples due to being snapshots of cell insights or patients. Intrinsically, a given list of genes from omics does not cover all associated genes. Moreover, insignificant but critical genes mediating biological processes and linking omics hits would not be specified due to the lack of alteration in cell metabolism. The identification of these essential genes, and downstream analyses of network-based methods strongly rely on the coverage of reference networks. Thus, my initial study focuses on the performance of network inference algorithms across different reference networks.

The reference networks are mainly composed of integrated databases. Due to high-throughput methods, the accelerated accumulation of knowledge enhances the complexity of references. Still, a simple network approach like this might guess the wrong genes and miss interactions that are specific to the situation because there aren't many omics hits in these complicated interactomes. Over-research on cancer comes with high annotations and a bias in prior databases. Thus, well-studied proteins prevent network-based methods from accurately identifying altered signaling networks and context-specific interactions due to the high number of interactions. The second motivation was to reduce the complexity of reference networks before network inference by avoiding biased information.

1.3. Contributions of the study

Beyond the list of molecules, it is essential to comprehend the molecular interaction with the collective consideration of multi-omics, elucidating rewiring and perturbations in cellular signaling cascades. The main question is how to accurately integrate multi-omics datasets with precise molecular interactions in the reference

networks. Thus, as an initial point of this thesis, we compared the outstanding human interactomes composed of experimentally known PPIs based on confidence scores, pathways, cancer driver proteins, structural information of protein interactions, and the bias toward well-studied proteins. The coverage affects the performance of network inference methods and influences downstream analyses and their interpretations. Thus, this thesis can guide researchers in choosing a reference interactome with their consideration of follow-up analyses, including structural evaluations of interactions and signaling alterations on pathways.

In this thesis, we developed pyPARAGON (PAgeRAnk-flux on Graphlet-guided network for multi-Omics data integratioN), a novel tool that merges network propagation with graphlets. pyPARAGON is available at **https://github.com/metunetlab/pyPARAGON**. Due to the sparse data, network features (e.g., degree distribution, clustering coefficients) have limited usage in propagation or inference. Committing knowledge of graphlets and their statistics, such as graphlet degree distribution, graphlet degree vector, and probabilistic approaches to graphlets, requires high computational force in complex networks due to the multiple interactions of nodes. Pioneeringly, we systematically recruited graphlets composed of known inputs and intermediate genes/proteins that link known molecules, which allows a computational advantage by targeting only associated regions. pyPARAGON identifies a core region composed of more functionally associated molecules and their interactions by trimming many interactions and decreasing the number of highly connected proteins in reference networks. Additionally, interactions are scored by normalizing with the total interaction count of connecting molecules, which penalizes highly connected proteins. Thus, pyPARAGON significantly overcame irrelevant interactions and proteins and outperformed other prominent network reconstruction tools.

In clinical applications, pyPARAGON effectively inferred various context-specific networks, including disease network models for complex diseases and patient-specific networks for 105 breast cancer patients. Recent epidemiological research on large cohorts of patients with autism spectrum disorder (ASD) revealed a higher cancer risk than in the general population. In the network models of cancer and ASD, we identified

their signal alteration on common pathways through different transcription factors by comparing the transcription profiles of pathway components. These outcomes reflect cell cycle effects—proliferation and differentiation—in distinct fates. A stronger signal level in their shared pathways promotes proliferation in cancer than differentiation, while mild signal levels enhance differentiation. This analysis critically examines the signaling strength of pathways in cancer and ASD, mentioning their differences and commonalities. The other case study with pyPARAGON identified meaningful network modules in patient-specific networks. The survival of patient clusters significantly differs based on functional modules, demonstrating altered pathways in their disease models. We frequently saw modules associated with the Ras signaling pathway as dominant in the patient cluster with the lowest survival probability. Beyond network inference, pyPARAGON provides additional meaningful downstream analysis for disease models, easing biological interpretation for researchers. The modular usage allows researchers to integrate pyPARAGON with various kinds of outputs.

1.4. Organization of the dissertation

The thesis has been published with six main chapters, titled “Introduction,” “Literature Review,” “Performance Assessment of the Network Reconstruction Approaches on Various Interactomes,” “pyPARAGON: Combining Network Propagation with Graphlets to Integrate Multi-Omics Data,” “Implementation of pyPARAGON” and “Discussion.” In Chapter 1, this thesis briefly describes the main concept explanations, motivation, and contributions of this study.

Chapter 2 explains the theoretical background of contextualization, covering recent approaches, such as learning- and network-based methods. After discussing generic biological networks, the review focuses on multi-omics data integration methods within the contextualization concept. Next, we detail network-based contextualization and its biological interpretations using community detection methods, overrepresentation, and enrichment analyses. At the end of Chapter 2, graphlets are described with graphlet-based metrics.

Chapter 3 comparatively assesses network reconstruction approaches such as all-pair shortest path, heat diffusion, personalized PageRank, flux calculations, and prize-collecting Steiner forests using various interactome as a reference network. Furthermore, biases, degree distribution, and coverage of prior knowledge, such as pathways and structural information, are used to evaluate reference interactomes. In this way, the thesis addresses challenging issues in network reconstruction algorithms and reference networks.

Chapter 4 introduces our novel tool, pyPARAGON, and its algorithm. The performance of pyPARAGON is compared with other tools by reconstructing cancer-associated signaling pathways and modeling different cancer types.

Chapter 5 exemplifies the use of pyPARAGON. Phosphoproteomics datasets are contextualized in a supervised manner to infer tumor-specific networks in the first case study. In the second case study, contextualized network-based disease models are the unsupervised usage for the comparison of two distinct but associated diseases, autism spectrum disorder and breast cancer.

Chapter 6 discusses the challenging issues in network-based approaches and how pyPARAGON solves these issues. The strengths and weaknesses of pyPARAGON are clarified by mentioning how to trim a reference network and reconstruct signaling pathways successfully. In addition, this chapter also discusses case studies and their outcomes.

CHAPTER 2

LITERATURE REVIEW

This chapter briefly overviews the background literature and relevant studies about biological networks and their interpretations. The literature review is discussed in three primary sections, namely: (1) biological networks, (2) contextualization of biological networks, and (3) graphlets.

2.1. Biological networks

The system biology approach employs biological networks to ascertain the interconnections among various data types, such as genomic, transcriptomic, proteomic, and metabolic[44]. Beyond the list of molecules, these networks illustrate molecular traffic in context, which models the time-dependent evolution of cells and other biological systems[5,6]. In biological networks, edges/links represent various interactions, such as protein-protein interactions (PPIs), transcription factors (TFs)-gene interactions, and small molecules, proteins, or gene interactions. The components of interaction are named nodes/vertices. By interpreting heterogeneous "omics" datasets, network approaches enhance our comprehension of perturbation mechanisms, therapeutic impacts, and gene functionalities within particular biological systems[45,46].

When examining how interactions occur, biological networks are classified into two distinct categories: directed networks and undirected networks. The former possesses edges that transmit sequential information, while the latter lacks such information on edges. Sequential database interactions, including pathways and signaling cascades, facilitate the anticipation of the corresponding segment of the information flow. Incomplete data in these databases, however, is one of the obstacles to inferring directed networks[47,48]. Networks can also be categorized as weighted or unweighted. Weighted networks allocate numerical values to the edges of the nodes, which

correspond to a range of characteristics, including experimental dependability, strength, cost, flow, and probability[49,50]. Unweighted networks, conversely, indicate the existence or non-existence of interactions between node pairs without assigning any score.

During topological characterization, interactions are examined across the entire network and in specific local regions to discover global and local characteristics, respectively. The characteristics of a network comprise clustering coefficients, degree distribution, and shortest path lengths[51]. The degree of a node indicates the number of interactions with other nodes. The minimum number of edges that must be traversed to travel from one node to another determines the shortest path between two nodes. Numerous network centrality metrics include, but are not limited to degree centrality, betweenness centrality, proximity centrality, and eigenvector centrality. These methods consider the significance of each node based on distinct criteria[52].

The nature of the multi-omics datasets between biological information establishes various biological networks, such as gene regulator networks, PPI networks, and cellular signaling networks. Gene regulatory networks (GRNs) include all of these interactions in specific contexts , e.g., tissues[53,54], drug treatment[46,55,56], mutations[57], gene knockouts[58,59], and disease conditions[60]. Since regulation occurs at each molecular level in cellular information flow, GRNs may include all other categories of biological networks in different contexts. Reference databases like JASPAR[61], iRegulon[62], BioGRID[63], TRRUST[64] and RegulonDB[65] integrate regulatory information. GRNs are directed networks since regulatory elements target specific genes/proteins at different levels[18]. Biological phenomena, such as catalysis, transportation, signal transduction, and growth control, occur with PPIs and are demonstrated with PPI networks. Proteins may have multiple functions due to various kinds of possible interactions. The functions of proteins depend on the time, location, and condition of PPIs[66,67]. On the other hand, the reference networks (e.g., STRING[50], HIPPIE[8], and iRefWeb[39]) just tabulate PPIs considering different criteria such as probability, the number of publications, or experimental evidence[9,50]. Thus, context-specific networks are essential to determine the associated functions of proteins. Cellular signaling networks integrate biochemical and physiological processes that

take place within the cell environment and occur sequentially to maintain cellular mechanisms necessary for the current situation of the cell, to activate or inhibit signaling cascades, or for the response of different internal and/or external signals[68,69]. Although reference databases (KEGG[70] and Reactome[71]) provide separate information on individual biological processes, each of which crosstalks via common proteins and PPIs inside the cell. Therefore, the inference of cell signaling networks can elucidate the specific interactions between various entities by simplifying complex crosstalking among signaling cascades[72].

2.2. Contextualization of biological networks

The contextualization of networks, using different databases, reveals the association between shared interactions and symptoms, during the analysis of interactions in diseases[73,74]. However, concerning complex cellular processes, analogies or correlations can be inadequate in comprehending causal relationships due to rewiring in networks[75,76]. However, context-specific solutions simplify a vast database into a small system, focusing on specific biological contexts such as rare diseases[77], specific tissues, or localized perturbations[78]. For example, the Parkinson's disease (PD)-specific network identified the associated genes and pathways and revealed a distinct expression pattern in differentially expressed drug-target genes[79]. Another study contextualized single-cell RNA sequencing (scRNA-seq) at multiple time points to understand reprogramming mechanisms in cellular conversion and transcriptomics states during lineage differentiation from embryonic fibroblasts to neuronal cells[80].

Many ontology databases exist for annotations, such as the Gene Ontology (GO)[81], ENCODE[82], Human Disease Ontology (DO)[83], and DisGeNET[84]. These databases are often valuable for adding details about biological processes, molecular activities, and cellular components to gene sets and diseases. They include single or multiple functions of genes, independent of biological contexts. The annotations from Gene Set Enrichment Analysis (GSEA) allow for identifying biologically significant processes enriched in specific contexts[85–87]. Beyond the gene sets, gene networks provide functional interactions, including activation, repression, or phosphorylation[88]. Genes/proteins exhibit varying roles contingent upon contextual factors, such as the specific cell type[89], drug introduction[75], or stress response[90]. Particular genes can

encode moonlighting proteins, which exhibit diverse functionalities due to the different isoforms and interactions[91]. For instance, EGFR impacts transcription, signal transduction, cell division, survival, motility, and other biological processes[92]. The challenging part is picking these genes to study further and figuring out what those roles are in specific biological settings. Therefore, researchers often rely on their experience and comprehensive literature searches to determine the relevant gene activities. Although acquiring their specific knowledge is valuable, it may be time-consuming and not always feasible in uncharted biological environments. Thus, network-based approaches enhance biological insights by integrating omics datasets and recruiting graph theory.

2.2.1. Data integration

The advancement of high-throughput omics technologies has facilitated the rapid accumulation of "big data" in biological and health sciences[93,94]. Using an integrative approach enables the elucidation of how various omics components are connected and how they affect each other[95]. The conceptual overview of multi-omics data integration is demonstrated in **Figure 1**. Significant biomolecules in omics can interact intimately and tightly regulate each other both inside and across various kinds of data. Improper interactions can potentially change cellular networks, eventually leading to aberrant signaling output. Thus, multi-omics data integration plays a pivotal role in comprehensively understanding the onset and development of diseases[96].

Prior knowledge in reference networks enables selecting related information and figuring out connections among omics entities during data integration. Several databases, including the Human Proteome Atlas[12], GTEx[97], and ENCODE[82], have annotated genetic variations and protein/gene expression patterns in different tissues. Additionally, multi-omics data for the same tumors, patients, or perturbations are combined in databases to investigate disease etiology. Multiple layers of omics data from thousands of tumor tissues in human cancers have been compiled by the Cancer Genome Atlas (TCGA)[23], the International Cancer Genome Consortium (ICGC)[24], Clinical Proteomic Tumor Analysis Consortium (CPTAC)[25], and TARGET[26] for pediatric cancers. The Cancer Cell Line Encyclopedia[27] and Cancer Dependency Map (DepMap)[98] store genomic and transcriptomic information, genetic dependency, and

slight chemical sensitivities of cancer cell lines for therapeutic response. Recently, patient- or condition-specific multi-omics data storage and their integrative analysis-based therapy techniques are moving quickly.

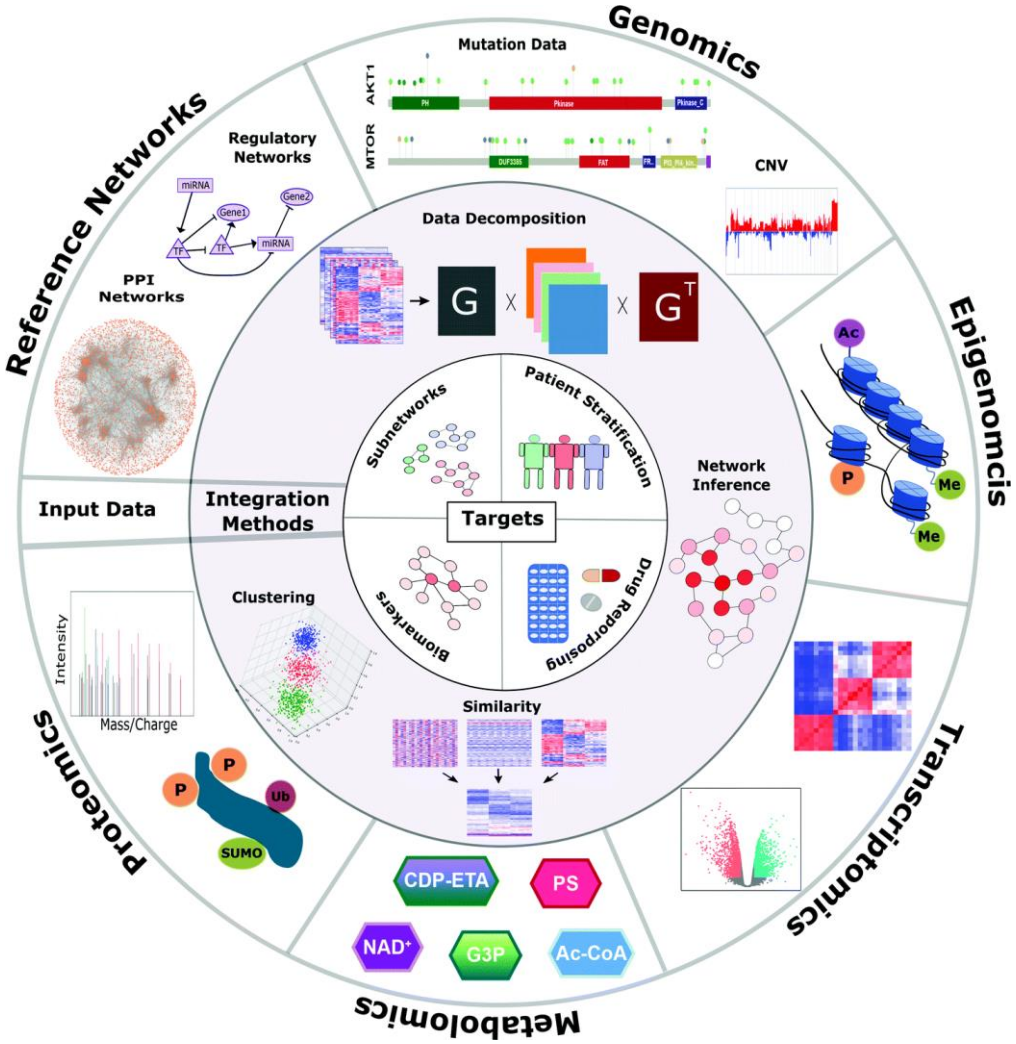

Figure 1: The conceptual overview of multi-omics data integration approaches and their applications. From the outer layer to the inner, the input omics data types, the integration methods, and their applications, respectively. High-throughput multi-omics data includes genomic, epigenomic, proteomic, post-translational modifications, metabolomic, and transcriptomic datasets. Depending on the method, these data may be integrated with or without a reference interactome. The inner cell interaction network carries information on different levels. A reference interactome may contain protein–protein interactions, regulatory interactions, metabolite-protein interactions, or others. As shown in the middle, network-based machine learning and statistical methods or their combinations can be employed for data integration. The innermost circle illustrates the final aim of integration tools, such as subnetwork construction, biomarker identification, patient stratification, and drug repurposing.

The fundamental problem in multi-omics data integration is creating effective and practical tools for reverse engineering big data to explain the molecular basis of a disease or a disturbance[99,100]. Integration methods can be categorized into two ways, namely horizontal and vertical, depending on the kind of datasets[101,102]. Horizontal integration involves using identical data types obtained from numerous samples, such as the aggregation of transcriptome data derived from various patients. In contrast, integrating several layers of omics data is used vertically, e.g., the connection between gene expression and mutation profiles. An instance of horizontal integration may be seen in the use of hierarchical HotNet to analyze pan-cancer somatic mutation patterns, leading to the identification of cancer-driver subnetworks[43]. In contrast, iCell employs a vertical integration approach that leverages tissue-specific protein-protein interaction, gene co-expression, and gene interaction networks to identify rewired genes inside the network. These rewired genes are promising cancer biomarkers[103].

Integration techniques can also be categorized according to the sequence of data usage, namely sequential and simultaneous integration procedures (**Figure 2**)[29]. The evaluation and optimization of omics datasets are conducted individually using sequential approaches[33,104]. In network-based analysis, each improved data independently is mapped into networks. Each succeeding phase refines the results of the preceding data by narrowing the search area and increasing the data size. However, this approach reduces sensitivity since weak or missing signals might contain valuable information[42]. Missing signals in multi-omics integration can arise due to either experimental factors such as instrument sensitivity, sampling factors, or insignificant critical points. TieDIE, for instance, combines mutations and differential gene expression profiles with the PPI-proximity test to locate the final subnetworks by using two heat diffusion processes in succession[17]. Initially, the heat is transferred to other genes in the directed reference interactome through diffusion from the significantly mutated genes. Next, a similar application is carried out by the reference interactome received in the opposite direction. Finally, TieDIE aggregates results from both directions to construct the ultimate subnetwork. The curse of dimensionality arises in multi-omics studies when the dimensionality of the data grows, leading to an increase in data sparsity. In addition, integration approaches are susceptible to overfitting in learning-based methods, particularly when fitting supervised models, due to the high-

dimensional data structure[105,106]. Overfitting is a challenge for both sequential and simultaneous processes. Sequential approaches independently reduce dimensionality on each omics dataset to address this issue[4] (**Figure 2A**). Simultaneous integration approaches, on the other hand, handle all features at the same time (**Figure 2B**). For dimension reduction, they frequently employ learning-based methodologies such as component analyses (MOSClip)[107], non-negative matrix factorization[108], multi-variate analysis (mixOmics)[109], and Bayesian framework (iClusterBayes)[110]. These strategies can address biases in multi-omics data, mitigating the risk of information loss[111,112].

Figure 2: Integrative network-based approaches. **A)** In the sequential integration approaches, some integration methods separately map an initial node-set (red and blue) from each omics data on the reference networks. However, the lack of direct connections of initial node sets causes incomplete subnetworks in integrated omics data. Network propagation methods such as random walks, heat diffusion, and the prize-collecting Steiner tree identify the hidden nodes (green) and construct subnetworks. **B)** In the simultaneous integration approaches, some tools directly integrate multi-omics data using statistical- or learning-based methods such as principal component analysis, joint multivariate regression, nearest shrunken centroid, or joint similarity matrix regardless of reference networks and primarily for identification of essential nodes (orange). Then, these nodes are leveraged to identify a subnetwork.

2.2.2. Contextualization approaches

Contextualization of biological networks recruits learning-based and network-based approaches to uncover biological relationships among omics entities.

2.2.2.1. Learning-based approaches

Depending on context awareness, learning-based methods are mainly intended to extract biological insights from large multi-omics datasets for classification[113], clustering[114], and ranking[115]. Their applications provide automated models from large multi-omics datasets to leverage interactions across omics layers. Fundamentally, learning-based methods are classified into supervised and unsupervised learning methods.

Supervised learning approaches prioritize predictions in context by determining distinctive rules from the data. They train models using labeled datasets to get the desired labels, such as cancer driver genes, disease-associated pathways, or drug responses[95]. Model training with large multi-omics datasets can be time-consuming[116]. Moreover, complex diseases and biological problems do not have specific boundaries in datasets. One example is CapsNetMMD, a supervised deep learning method that uses multi-omics data as input to a two-layer convolutional neural network to sort genes that are linked to breast cancer[117]. Another tool is DeepDRK, which utilizes multi-omics datasets derived from various drug-treated cell lines and drug characteristics to make predictions on cell line drug sensitivity[19]. Another recent deep learning tool using convolutional neural networks, MOGONET, enables the classification of patients and the identification of biomarkers[113]. In the reverse engineering approach, using single-cell RNA-sequencing (scRNA-seq) and single-cell assays for transposase-accessible chromatin using sequencing (scATAC-seq) data, linear regression models identified expressed genes and the regulatory mechanism by simulating biological insights of a single-cell[118]. In the other case, training linear regression models with large reference panels like GTEx lack gene expression weights, estimated in specific contexts[119,120]. To achieve optimal regression and identify significant variables, the model requires ample sample space and the implementation of grid search. Hence, the use of direct regression for statistical inference comes with computing challenges due to the high dimensionality of the data[95].

Unsupervised learning approaches attempt to discover latent structures or patterns in unlabeled datasets. Due to the lack of ground truth datasets, heuristic methods like clustering quality metrics are mainly used for method assessments, which might allow for biased evaluations[121]. Various multi-omics integration technologies use similarity metrics, kernels, and statistical methods to construct unsupervised learning techniques. Similarity-based integration is a prevalent approach in addressing patient stratification, considering the distances in the multi-omics data across individuals. Similarity Network Fusion[122] and Perturbation Clustering for Data Integration and Disease Subtyping[123] are mainly recruited for similarity-based integration in biological contexts. For versatile integration of multi-omics datasets, rMKL-LPP[124] and MixKernel[125] use different kernel learning methods. The rMKL-LPP method implements the Locality Preserving Projections (LPP) algorithm to reduce the dimensionality of the data while maintaining the similarities and closest neighbors.

Statistical methods can establish models that capture the associations between characteristics exhibiting the most significant biological variability. These models recruit various methods, such as correlation techniques, regression formulas, and assumptions about probability distributions. Most of the contemporary tools can incorporate diverse data types, including binary (representing somatic mutation), categorical (representing copy number gain, regular, or loss), and continuous (representing gene expression) data, each of which follows distinct probabilistic distributions[95]. However, specific tools, like iCluster[104] and JIVE[126], cannot simultaneously handle discrete and continuous data. Various statistical techniques, such as generalized principal component analysis, e.g., MOFA[127] and mixOmics[109], or low-rank approximation approaches, e.g., iClusterBayes[110] and LRAcluster[128], are used to deconstruct datasets to elucidate the underlying shared variation, individual variation, and noise.

2.2.2.2. Network-based approaches

Network-based approaches leverage graph theory to reveal the dependencies among omics entities in a given context. Usually, a reference interactome is mainly employed during contextualization, covering protein-protein interactions, gene co-expression, metabolite interactions, and regulatory interactions[129–131]. Network-based algorithms

depend on network features such as degree distribution and clustering coefficients. These approaches, including neighborhood-based and shortest-path algorithms, take into account the localized interactions of context-specific seeds/initial nodes that may be derived from omics datasets or other databases individually. In contrast, diffusion-based algorithms use the global characteristics of reference networks and systematically assess initial nodes (**Figure 3**)[132].

Neighborhood-based methods involve establishing connections between all pairs of seeds/ initial nodes—context-specific information that can come from omics datasets or other databases—based on reported interactions in a database. Many tools are available for expanding seed nodes using k-step interactors. Notable examples are STRING[50] and BIANA[133], which can be accessed via the Galaxy and InteractoMIX platform[134]. The outputs of these methods may include either a compact, unified network or several disjointed components, depending upon the initial nodes and their related interactions. In this approach, network inference is based on guilt by association, whereby entities interacting with each other are likely to be part of comparable functional modules and contribute to the same biological processes[135]. Each component of a given protein complex is always connected to at least one other component, establishing an intimate proximity relationship. Nevertheless, highly connected reference networks might produce false interactions using neighborhood-based approaches in contextualized networks[136].

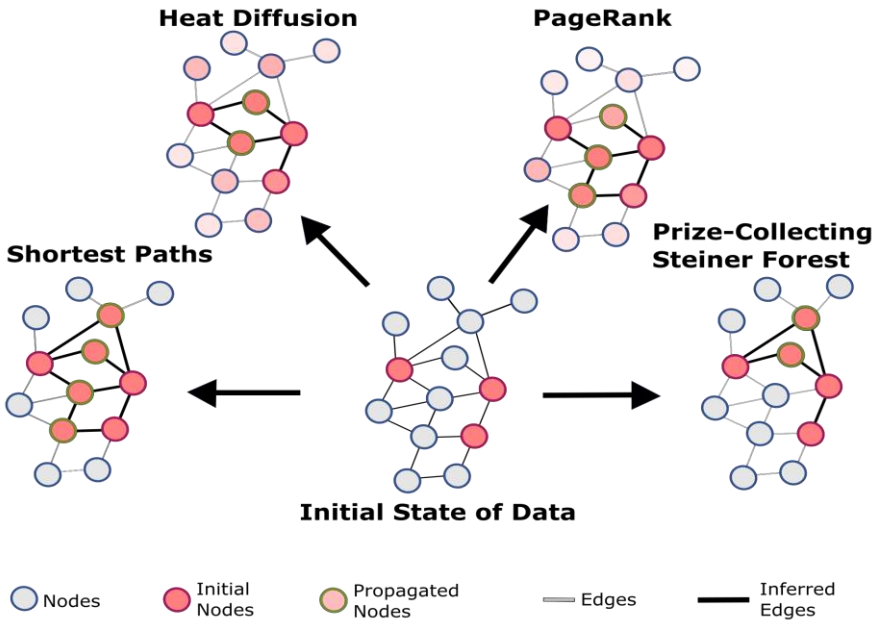

Figure 3: The conceptual representation of reconstruction algorithms. Shortest paths, personalized PageRank, heat diffusion, and the prize-collecting Steiner Forest. The shortest paths reconstruct the subnetwork by combining all or highly scored shortest paths between seed nodes. Heat diffusion splits heat that belongs to initial nodes. Heat diffusion splits heat that belongs to the initial nodes. After limited steps of transfer, the heat of the nodes can be used for edge scoring. PageRank uses the probabilities of nodes randomly walking in the reference interactome. The prize-collecting Steiner Forest application finds the optimum forest to link seed nodes directly or through intermediate nodes. The union of optimum forests reconstructs subnetworks.

Shortest path algorithms connect initial nodes by identifying the shortest paths for each initial node pair in generic reference networks and appending related nodes and interactions along the shortest path. PathLinker connected initial nodes by scoring the k-shortest paths and ordering nodes from receptors to transcriptional regulators[137]. Lists2Networks, a web-based system, integrates co-expression or background knowledge co-annotation correlation by applying gene-list enrichment analyses against prior biological knowledge, such as pathways and gene ontology terms[138]. POINeT simplified the subnetwork construction process by combining PPIs and tissue-specific expression data from multiple resources, filtering peripheral nodes, and assessing the confidence scores of interactions. However, including all possible paths would result in an expansion of the network size, longer computation time, and an increase in false positive results. Using approximation approaches can save computational time; nonetheless, optimizing large networks might be demanding regarding resources[139].

Network diffusion methods are commonly used in modeling molecular data because they are flexible, easy to represent, and can find complex patterns and clusters by propagating initial nodes in reference networks. After a certain number of iterations, or until convergence, prior information from seed nodes disperses neighbor nodes through edges. In each iteration, all nodes inside a network are impacted by their adjacent nodes, and so on in the following iterations[132]. Ultimately, new nodes, not among initial nodes but associated with the given context, are combined with prior knowledge by constructing their interactions. Random walks, PageRank, heat diffusion, and belief propagation algorithms are examples of network diffusion algorithms[38]. By applying the random walk algorithm to the TCGA dataset, MEXCOwalk finds modules comprising known and probable cancer genes based on somatic mutations[35]. Implementing the random walk algorithm, uKIN merged known disease genes and new putative disease genes to find functionally relevant genes for several complex disorder[36]. As an example of heat diffusion algorithms, HotNet2 revealed cancer-associated signaling pathways and subnetworks with rare somatic mutations across multiple cancers using TCGA data[37]. Another application of heat diffusion, TieDIE, utilized a compilation of frequently mutated breast cancer genes from the Catalogue of Somatic Mutations in Cancer and collections from The Cancer Genome Atlas (TCGA) to provide a mechanistic understanding of tumor characteristics and propose subtype-specific drug targets[17]. Several methods in network reconstruction address the Steiner tree issue by creating a tree with the lowest possible cost involving seed nodes. Omics Integrator 2 employed the prize-collecting Steiner Forest approach to construct network models of complex diseases, intending to determine shared and distinct pathways[140]. Furthermore, Omics Integrator 2 identified potential drugs specific to each disease by stratifying patients[141] and attainable combinations of drugs[142].

Both the neighborhood and shortest-path strategies are local methods, which means they focus on analyzing and making predictions based on immediate connections or relationships. However, it is essential to note that these methods may not effectively capture or account for perturbations that arise from peripheral connections or relationships. This phenomenon may be illustrated when initial nodes establish connections with more significant components via a limited number of linkages[132].

The presence of various, equal, or multiple shortest paths between seed nodes may lead to subjective methods[137]. Network diffusion-based algorithms might overcome these constraints. Each node is impacted by the entire topology of the reference network through an iterative process that involves information exchange between neighboring nodes. Also, each node is allocated a quantitative value, determined by its proximity to multiple seed nodes and the global topology of the network[95]. When establishing reference networks that include integrated interactions or relationships, disregarding the context, there might be false positive and negative interactions in context-specific networks. One of the critical reasons, well-studied proteins with hundreds of high-confidence interactions, also known as hub proteins, like TP53 and EGFR, lead to a bias in the reference networks[143]. Some network-based methods, like Omics Integrator[34], TieDIE[17], and Hierarchical HotNet[43], use context-specific interactions and punish the hub nodes to overcome this bias.

2.2.3. Interpretations of context-specific networks

Context-specific networks may be annotated by examining the intersecting members, including nodes and edges, obtained from various databases such as pathways, biological processes, regulatory elements-targets, disease databases, etc. Despite being a reduced form of complex large networks, the assessments of all nodes in the context-specific networks provide extensive and unspecific annotations due to over-crosstalking among annotations. Thus, small interacting communities or clusters in networks can be more informative than overall inferred networks[144]. Overrepresentation and enrichment analysis are standard methods for interpretations to reveal significant communities associated with a given serial process, such as a reaction, a pathway, or a biological process. Overrepresentation analysis determines whether a previously isolated specific community significantly associates with pathways or cascades. On the other hand, enrichment analysis examines differential data from all measured nodes and identifies processes demonstrating significantly coordinated shifts, such as activated or inhibited[145]. As well as biological annotations, the use of prior knowledge in databases can refine the inferred networks by pruning insignificant members or expanding the skipped members within a particular context. For example, if a cluster of proteins in the network forms a part of a known protein complex, the remaining members of the complex could be added (both nodes and

edges). Also, insignificantly identified clusters in context-specific networks may be subjected to additional filtration using permutations on datasets or reference databases to eliminate noisy nodes and interactions[132].

2.2.3.1. Community detections

Community detection has arisen within network research to identify clusters inside complex networks. Conventional decomposition techniques aim to identify a rigid block diagonal or block triangular structure. On the other hand, community detection approaches focus on identifying subnetworks with statistically more connections between nodes within the same group than those across various groups[146]. A benefit of employing community detection to find decompositions is that the subproblems, established in context, will have statistically minimum interactions via complicating variables or restrictions and demand little balance through the decomposition solution approach. Thus, community detection methods provide evidence for dynamic views of modules in context-specific networks where different groups of nodes perform distinct functions[147,148].

Communities can be regarded as state variables that exhibit densely interconnected interactions among their components while demonstrating relatively weaker interactions with other communities. The identification of communities requires the optimization of the quality function known as modularity, which is a measure of connections within a community[149]. Agglomerative and divisive techniques are the two primary categories that may be used to classify different community identification methods considerably[150]. Agglomerative techniques include adding edges to a graph that initially consists only of nodes. Edges are appended in a unidirectional manner, from the more robust edge to the more vulnerable one. Divisive approaches follow the opposite of agglomerative procedures. In this scenario, the process involves gradually removing edges from a complete graph. In a particular network, there are a variable number of communities, each characterized by its own size. The qualities of context-specific networks provide significant challenges to community detection. Followingly, the most commonly used community detection strategies are explained below.

The Louvain algorithm was introduced in 2008 as a heuristic approach for efficiently identifying communities in large networks[151]. Based on modularity, this method seeks

to maximize the deviation between the observed and predicted edge counts in each community. The Louvain method is structured as a repetitive process consisting of two distinct phases: the local movement of nodes and network aggregation. During the former stage, the algorithm assigns a specific community to every node inside the network. Here, the algorithm aims to reach a local maximum of modularity. In the latter, the algorithm constructs a new network by treating the communities identified in the previous phase as individual nodes. The neighbors of each node are examined by calculating the change in modularity that would result from removing the node from its present community and putting it in one of its neighboring communities. The node will be positioned inside the neighboring community if the gain is positive and maximized. The node will continue to reside within the same community if no positive benefit exists. This iterative process is implemented for all nodes until no more enhancements exist. The popularity of the algorithm stems from its effortless implementation and impressive computational efficiency. Nevertheless, a prominent constraint of the algorithm resides in its reliance on the storage of the network within the primary memory[152].

Recent research in 2019 by Traag et al. demonstrated that Louvain community detection has the propensity to identify internally disconnected communities, also known as weakly connected communities[153]. However, strong connections between other nodes in the community enable it to maintain its status as a distinct community. In the Louvain algorithm, changing a node that connects two parts of a community to a different community might break the connection between the two parts of the old community[152,153]. Therefore, crosstalking between biological processes may be disrupted in context-specific networks. To enhance the quality of the identified partitions, the Leiden algorithm ensures robust interconnectivity across communities through an additional step between two phases of the Louvain algorithm. Here, the communities in the first step can be further divided into many divisions in the second phase. As part of the refining process, a node may join a randomly picked community to make the quality function higher. This randomization allows for a broader discovery of the partition space.

The recent approach, Surprise, a statistical measure of interest based on classical probabilities, assesses the quality of a network partition into communities due to modularity constraints. The assumption of random interaction between nodes in a network underlies the idea of Surprise. Based on the hypergeometric distribution, the method finds a possible division different from how the community nodes and connections should be spread out[154]. The use of Surprise can be effective in detecting a large number of small communities. In contrast, the implementation of modularity is advantageous in the identification of a limited number of communities.

2.2.3.2. Overrepresentation and enrichment analysis

The identification of communities inside context-specific networks allows for the analysis of topological interactions among nodes and communities. Overrepresentation analysis (ORA) can annotate communities with biological processes, pathways, signaling cascades, and other relevant elements to understand the biological functions and underlying pathological phenotypes. ORA facilitates the identification of motifs within communities, thereby enabling the successful elucidation of molecular machines within context-specific networks. The statistical test most often used in ORA relies on the hypergeometric distribution or binomial approximation when evaluating communities or genes/protein sets[155]. Conventional techniques for ORA use the most significant hits from datasets, considering several factors, such as fold changes and the significance of changes. During evaluations, ORA tools examine each component in a community independently, without considering the given weights for seed nodes. Consequently, certain relevant information may get obscured or omitted. Conventional iterative approaches append one or a few genes at a time to extend a set of genes[156,157]. However, network-based approaches to contextualization can propagate the initial information and score components of communities for ORA[158].

Enrichment analysis (EA) tools, unlike ORA, utilize the weights of initial nodes from omics datasets or the scores of components in context-specific networks. EA Tools provide directional representation such as up- and down-regulations on pathways or enriched and depleted components, evaluating the impact of communities by statistically comparing functional information such as pathways and biological

processes[155]. Most EA tools utilize univariate and multivariate methods. After calculating gene scores, univariate methods utilize straightforward statistics or distributions, such as calculating the mean of the squared t-statistics of genes inside the gene set, using either sample or gene randomization. Here, gene randomization might assess the statistical significance of the results[159]. Multivariate methods go straight to the computation of gene set scores using the expression matrix without the intermediate step of computing gene scores. Multivariate tests may provide statistical power due to their ability to include interdependencies across genes by evaluating the joint distribution of gene expression levels[160].

2.3. Graphlets

Incorporating subgraph information in network interpretations leverages the comprehension of complex cellular networks. Graphlets, small connected and non-isomorphic subgraphs, provide valuable insights into biological processes through neighborhood-oriented assessments of several nodes[47,161].

2.3.1. Graphlet-based metrics

Graphlet statistics systematically evaluate node placement within graphlets instead of interactions between pairs of nodes by generalizing notions of network features.

The degree distribution quantifies the count of nodes with a particular degree value (k), 'touching' k edges. In fact, in the degree distribution, an edge, G_0, is only considered (**Figure 4**). Therefore, the degree distribution measures how many nodes interact with a single G_0. As an extended version, graphlet distribution quantifies how many nodes are in contact with a given graphlet. However, it is crucial to consider the position of nodes, called orbits, inside graphlets. For instance, the positioning of nodes inside G_1 might occur at either the endpoints or in the center of the graphlet. The concept of automorphic orbits refers to the isomorphic representation of orbits inside an individual graphlet. Hence, it is possible to see multiple automorphic orbits exhibiting identical topological features. The graphlet degree distribution (GDD) demonstrates the distribution of distinct nodes residing inside a certain graphlet. **Figure 4** displays 14 orbits belonging to 8 graphlets composed of 2, 3, and 4 nodes. The degree distribution, widely recognized as a global network attribute, belongs to a

set of 14 GDD that quantify the local structural characteristics of a network. It is essential to acknowledge that the GDD metric primarily focuses on assessing the local structure of a network since it is derived by analyzing tiny local network neighborhoods[162].

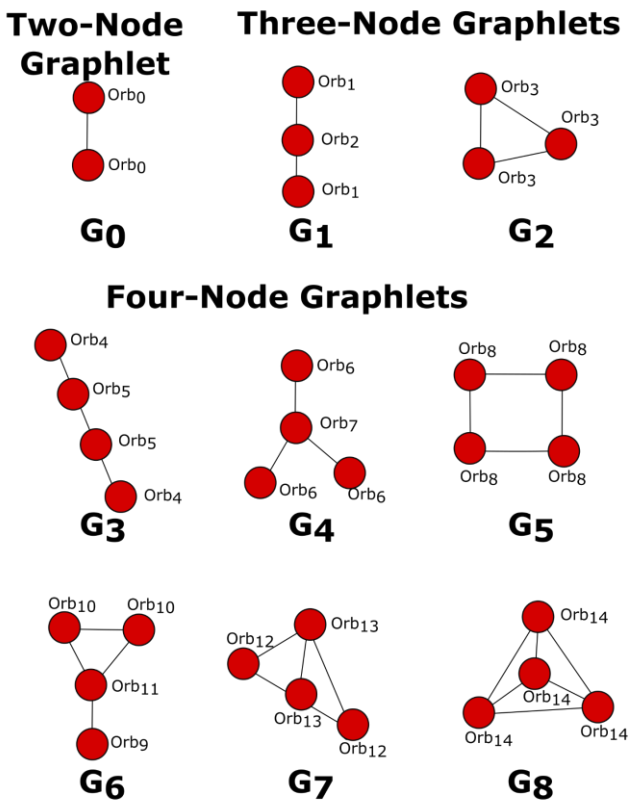

Figure 4: 14 Automorphism orbits for nine graphlets. 2, 3, and 4-node graphlets G_0, G_1, ..., G_8.

The Laplacian matrices, derived from the interactions between pairs of nodes, are utilized in spectral clustering, spectral embedding, and network diffusion to handle network issues in a computationally feasible manner. Generalized by defining a pair of nodes as 'adjacent' in pre-specified graphlets if they both interact with a certain graphlet, the Graphlet Laplacian Matrix includes graphlet-based topological information and node membership inside the same network neighborhood[103]. Graphlet degree vector (GDV) of a specific node demonstrates the number of a particular graphlet where nodes reside inside a particular automorphism orbit[161]. Node A in the dummy network is considered in **Figure 5** to show how graphlets may quantify the local topology surrounding a node. There are two instances of a graphlet G_1 where

nodes participate in three-node paths, A-B-C and A-B-E. The local network topology for a specific node may be quantified by generating a vector and calculating counts for that node over all graphlets for node a. Two nodes, u and v, in graph G are said to be graphlet-adjacent with regard to a particular graphlet in a given graphlet if they both reside in the graphlet. In the illustrated dummy network seen in **Figure 5**, nodes A and B exhibit graphlet-adjacency twice inside the network G_1. The graph-let-based adjacency matrix is defined in the following **Formula 2.1**:

$$A_k(u,v) = \begin{Bmatrix} a_{uv}^k \ if\ u \neq v \\ 0 \quad otherwise \end{Bmatrix} \quad (2.1)$$

where a_{uv}^k equals to the number of times nodes u and v are graphlet-adjacent in G_k. The concept of graphlet degree extends the notion of node degree by quantifying the number of times node u interacts with a certain graphlet G_k. The Graphlet Degree Matrix is derived from the degree matrix for a certain graphlet G_k, **Formula 2.2**:

$$D_k(u,v) = \begin{Bmatrix} d_{uv}^k \ if\ u = v \\ 0 \quad otherwise \end{Bmatrix} \quad (2.2)$$

where d_{uv}^k the number of times node u resides in graphlet G_k. Ultimately, the Graphlet Laplacian Matrix is defined in **Formula 2.3**:

$$L_k^G = D_k - (A_k / \theta) \quad (2.3)$$

Where θ = size (G_k) - 1. Going beyond the Laplacian Matrix only considering the neighbors, the Graphlet Laplacian Matrix, L_k^G, quantifies the strength of interactions between each node and all other nodes in the particular graphlet, G_k.

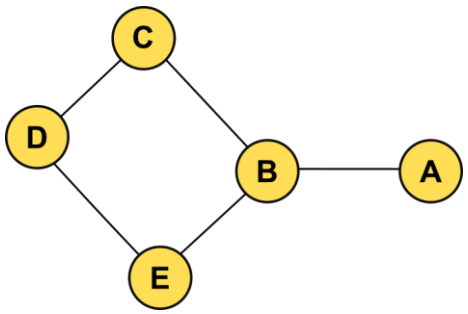

Figure 5: An example of a 5-node dummy graphlet.

2.3.2. Graphlet motifs

During network analysis, graphlets provide rich topological metrics by decomposing networks into small graphlet motifs that can be informative for functional subgraph patterns. For example, motif analysis identified the feed-forward loops in biological cascades, such as gene regulation and metabolic networks[163–165]. The graphlet motifs can be searched in static and temporal networks[166].

In static approaches, graphlet motifs can be defined as overrepresented graphlets in a given network through statistical methods. The number of graphlets in a given network is compared with randomized networks[167]. Several network models have randomized a large network, ranging from the initial Erdös–Rényi random graphs and geometric random graph models to more recent models such as small-world, scale-free, and hierarchical approaches. However, randomization of a network may disrupt local or global features that are informative for context-specific analysis. On the other hand, a geometric random graph model pertains to both features thanks to fitting a model to the similar degree distribution of a given network[168] In context-specific networks; permutation methods are also used as an additional method to enhance the robustness of interpretations. These approaches cover renaming nodes or modifying their weights, sustaining the distribution of node attributes. Furthermore, the initial nodes and their weights may be swapped during network reconstruction[169,170].

In temporal approaches, graphlet motifs are considered at a particular time to characterize networks. Temporal approaches do not directly capture situations where several events coincide[171]. Instead, they immediately add and remove edges within loosely connected subgraph edges, allowing for the analysis of networks[3,167]. Thus, these analyses depend on dynamic loss or gain in edges from one state to the next. This active transition knowledge inside a snapshot network offers the probability of a particular transition. For instance, graphlet patterns may be seen in various stages of gene network development across distinct locations, regardless of the roles of the individual genes[172].

CHAPTER 3

PERFORMANCE ASSESSMENT OF THE NETWORK RECONSTRUCTION APPROACHES ON VARIOUS INTERACTOMES

Beyond compiling molecular entities, it is essential to collectively analyze omics datasets and reconstruct molecular interactions to understand cellular mechanisms. Pathway reconstruction methods are important in comprehending disease biology, primarily due to the potential clinical consequences of aberrant cellular signaling. The primary obstacle lies in effectively combining the data in a precise manner. The objective of this section is to do a comparative examination of different network reconstruction approaches on many reference interactomes. Initially, various human interactomes were examined based on the coverage of each interactome concerning cancer driver proteins, the availability of structural information on protein interactions, and the potential bias towards well-researched proteins. Subsequently, the interactomes were used for the effectiveness of four outstanding network reconstruction approaches: all-pair shortest path[137], heat diffusion with flux, personalized PageRank with flux[173], and the prize-collecting Steiner Forest (PCSF)[34]. Each approach carries its own merits and limitations. We reconstructed curated cancer signaling pathways from NetPath, recruiting selected interactomes and reconstruction approaches. PCSF had the most balanced performance in terms of precision and recall scores. The successful implementation of each network reconstruction methodology is heavily contingent upon the quality and accuracy of the reference interactomes[38].

3.1. Methods

3.1.1. Reference interactomes

We utilized interactomes, including PathwayCommons v12[174], iRefWeb v13[40], HIPPIE v2.2 and v2.3[8], ConsensusPathDB v34[175], STRING v11[176], and OmniPath[177]. These interactomes integrate different and various types of protein-protein interactions (PPIs) databases covering pathways, biological processes, and experimentally

identified PPIs. The node and edge information of interactomes is detailed in **Table 1** after removing self-interactions and repeated interactions. PathwayCommons and OmniPath do not provide confidence scores to demonstrate the reliability of their interactions, while the other interactomes employ different confidence score schemes. The MI-scoring scheme utilized by iRefWeb considers several parameters, including experimental detection methods and scales of studies (low- or high-throughput). HIPPIE and ConsensusPathDB have confidence scores on edges calculated based on their own schemes. We strengthened STRING by only allowing experimentally validated PPIs to participate in the interactome.

Table 1: Statistics of interactomes

Interactome	Number of proteins	Number of interactions	Confidence score
iRefWeb v13	11,295	80,351	Yes
PathwayCommons v12	18,536	1,126,072	No
HIPPIE v2.2	15,984	369,584	Yes
ConsensusPathDB v34	17,269	359,201	Yes
STRING v11	8,922	229,306	Yes
OmniPath	6,549	35,684	No

3.1.2. Interactome comparison metrics

At both the node and edge levels, the reference interactomes were compared using the overlap coefficients for different metrics like the overlap coefficient, the correlation of edge confidence scores, the presence of proteins linked to diseases, and the coverage of pathway edges.

The overlap coefficient is a similarity metric comparing two datasets, S1 and S2. These datasets can represent node sets or edge sets derived from a database. The calculation of the overlap coefficient was performed using **Formula 3.1** to compare interactomes in pairs and assess the extent of knowledge coverage[178].

$$overlap(S_1, S_2) = \frac{|S_1 \cap S_2|}{min(|S_1|, |S_2|)} \quad (3.1)$$

We defined each pair of interactomes as G (V_G, E_G, $c(e_G)$) and H (V_H, E_H, $c(e_H)$), where V is the node set. E is the edge set, and $0 \leq c(e) \leq 1$, where c(e) is the confidence score of an edge. If the reference interactome does not have confidence scores, $c(e) = 1$ is uniformly defined for algorithms. The node-level similarities of the respective interactomes were determined by using the overlap coefficient, as described in

Formula 3.1; V_G and V_H are recruited as S_1 and S_1 in the given interactomes, G and H. Similarly, E_G and E_H are utilized as S_1 and S_1 in the calculation of edge-level similarities.

This thesis examined structurally known PPIs already established in reference networks. We retrieved structural knowledge from INSIDER[179], composed of 4,150 experimentally known interactions from the PDB[180] as well as 2,901 predicted interactions from Interactome3D[181]. We calculated the edge-level overlap coefficient between each reference interactome (G) and each part of INSIDER (H) through **Formula 3.1**.

Interactomes and network reconstruction approaches are often used to identify cancer driver modules. We retrieved the 568 cancer driver genes (CDGs) from intOGen[182]. Based on nodes, we calculated the overlap coefficient between CDGs (S1) and proteins in each reference interactome (S2). Moreover, we examined a bias towards cancer-associated proteins in reference networks with the number of publications about each CDG and the degree centrality of the CDGs, after retrieving PubMed IDs of 20,413 proteins from UniProtKB.

Pathway enrichment and overrepresentation analysis mainly explain the functionality of networks or subnetworks. The overlapping regions of a reference network (G) with 171 pathways established in KEGG (H) were assessed with the overlap coefficient calculation[183]. As a challenging issue, a small group of molecular interactions restrict the overall activity of signaling cascades in modeling small-sized networks[184–186]. Thus, we only evaluated pathways with more than 30 edges.

In the pool of interactomes, iRefWeb, HIPPIE, ConsensusPathDB, and STRING have edge confidence ratings using various scoring methodologies. We performed an all-pair comparison of the supplied interactomes (G, H) and a pearson correlation analysis on the confidence ratings in the intersection of edge sets ($E_G \cap E_H$).

Biological networks have a scale-free power law distribution, **Formula 3.2**, where k is a node's degree and γ is the power coefficient[187,188]. To obtain a linear representation of both the degree distribution and the publication distribution, the logarithm of the

distribution was recruited in **Formula 3.3**. The number of publications belonging to protein was retrieved from UniProt. A pearson correlation test on a logarithmic scale assessed the relationship considering the degree of nodes and the number of publications belonging to nodes.

$$P(k) = k^{-\gamma}\, P(k) = \; k^{-\gamma} \qquad (3.2)$$

$$log(P(k)) \; = \; -\gamma log(k) \quad (3.3)$$

3.1.3. Network reconstruction methods

This thesis evaluated four reconstruction approaches: the shortest path, heat diffusion, PageRank, and PCSF. The reference network, or a given interactome (G), was individually employed with node (V_G) and edge (E_G) sets and the weight of edges (c(e)). Network reconstruction methods aim to identify the subnetwork, defined as $R(V_R, E_R)$, where $V_R \subseteq V$ and $E_R \subseteq E$, by establishing connections between the seed nodes, $V_I \subseteq V$. The provided seed nodes were assigned uniform weights ($1/|V_I|$), where $|V_I|$ represents the number of seed nodes. The remaining nodes are assigned a weight of 0, allowing for the definition of w(v) in reconstruction methods.

3.1.3.1. All-pairs shortest paths

We determined all the most straightforward paths between each pair of nodes, u and v $\in V_I$ and $u \neq v$. In instances with several shortest paths connecting nodes u and v, we considered all such pathways. Ultimately, the integration of all paths created the final subnetwork. We did not apply any edge weight-based filtration or route length cutoff.

3.1.3.2. Personalized PageRank

The PageRank method was initially devised for propagation in directed graphs. Adapting personalized PageRank (PPR) to undirected graphs involves the conversion of each edge into two directed edges. The PageRank score, denoted as p(v), for each node in the reference interactome G measures the likelihood of being present at a particular node at a given time step (t). This probability is computed using the iterative **Formula 3.4**.

$$p_{t+1}(y) \; = \; \frac{1-\lambda}{N} \; + \; \lambda \sum_{x_i \to y} \frac{p_i(x_i)}{deg(x_i)} \quad (3.4)$$

where the probability of node y ∈ V is calculated using the damping factor (λ) that defines the probability of moving from neighboring nodes (x_i) to y, the total number of nodes in the interactome is referred to as N[189,190]. Initial probabilities of nodes were obtained from w(v). **Formula 3.4** was iterated 100 times as the default setting to get the probability distribution function p(v).

3.1.3.3. Heat diffusion

In the heat diffusion (HD) context, seed nodes with uniform heat distribution prioritize their associated nodes via heat transfer. This prioritization is mathematically defined in **Formula 3.5**.

$$p(v) = p_0\left(I + \frac{-\alpha}{N}L\right)^N \quad (3.5)$$

In **Formula 3.5,** L = I –W, where I denotes an identity matrix, and $W= D^{-1}A$, in which D and A represent the diagonal degree matrix and the adjacency matrix, respectively. The vector p represents the initial heat distribution in a system, where the nodes are assigned, weights based on the function w(v). The variables N and α represent the number of iterations and the heat diffusion rate, respectively. N is set to 3 as default [191]. After the heat diffusion process is completed, nodes have the diffused heat vector, p(v), as the weight.

3.1.3.4. Edge selection over flux scores

Personalized PageRank with flux (PRF) and heat diffusion with flux (HDF) is computed on the deg(v), p(v), and c(e), which represent the number of interactions, the probabilistic score obtained from PPR or HD, and the confidence score of a provided node in the interactome, respectively. In this thesis, unlike TieDie and HotNet, the threshold value was implemented to exclude uncritical nodes[173,192,193]. Also, during the subnetwork reconstruction, we considered the nodes in the interactome with $p(v_i) \geq 1/N$, where N is the number of nodes in the interactome. The directional flux scores ($f_{u\rightarrow t}$ and $f_{t\rightarrow u}$) were computed using **Formula 3.6** and **3.7**. The minimum of both directional flux scores (**Formula 3.8**) specified the final flux of the edge.

$$f_{u\to t}(u,t) = \frac{p(u).c(e)}{deg(u)} \quad (3.6)$$

$$f_{t\to u}(t,u) = \frac{p(t).c(e)}{deg(t)} \quad (3.7)$$

$$f(e) = min(f_{u\to t}(u,t), f_{t\to u}(t,u)) \quad (3.8)$$

The ranking of edges is determined by arranging them in descending order based on their flux scores, which are obtained by calculating the negative logarithm of the flux values. The calculation of the total flow (F) is performed by considering the interconnected nodes in **Formula 3.9**:

$$F = \sum f(e) \quad (3.9)$$

τ ($0 \leq \tau \leq 1$) is the scaling factor that is the threshold percentage of F. We selected the edges, from 1 to j, by summing flux scores up to τxF (**Formula 3.10**). The edges with low flux scores were eliminated from the reconstructed subnetworks[173].

$$\tau xF > \sum_{i=1}^{j} f(e_i), \quad 1 \leq j \leq n \quad (3.10)$$

3.1.3.5. Prize-Collecting Steiner Forest

We utilized Omics Integrator 2, implementing the PCSF approach as a state-of-the-art network reconstruction method. The costs of the edges are determined using the cost function provided in Omics Integrator 2 by combining the confidence score of the edge, c(e), with a penalty generated from the scaled node degrees using the parameter γ[34]. The updated version, Omics Integrator 2, applies a penalty to the edges according to the degrees of the node pair. The subsequent function aims to identify an optimal forest, denoted as F(V, E), by minimizing the objective function as described in **Formula 3.11**[194]

$$f'(F) = \sum \beta.p(v) + \sum cost(e) + \omega.\kappa \quad (3.11)$$

where κ represents the number of connected components; β is responsible for determining the relative weight of the node prizes; and ω influences the cost of adding a tree to the solution network. The PCSF algorithm yields an ideal forest for a given parameter set and an augmented forest that contains all edges connecting nodes inside

the optimal forest. The final networks were reconstructed by intersecting augmented forests using multiple parameter sets defined under parameter tuning.

3.1.4. Performance analysis

NetPath is a delicately curated collection of human signaling pathways, including immune and cancer signaling pathways. We used 32 pathways from NetPath as the benchmark dataset[195]. The numbers of nodes and edges are listed in **Appendix A**. Given the high computational cost associated with reconstructing all pathways using all parameter settings, it was necessary to first identify the optimal parameter sets before doing a performance comparison.

3.1.4.1. The Calculation of performance metrics

After adjusting parameter settings on four pathways, the remaining 28 NetPath pathways were implemented for the performance assessment with five-fold cross-validation. We tested each reconstruction method on its own on each reference interactome by calculating the F1 score, Matthew's correlation coefficient (MCC), recall and precision scores, and false positive rate (FPR) using **Formula 3.12** to **3.16**, as shown below[38].

$$recall(TP,TN) = \frac{|TP|}{|TP| + |FN|} \quad (3.12)$$

$$precision(TP,FN) = \frac{|TP|}{|TP| + |FP|} \quad (3.12)$$

$$FPR(TP,TN) = \frac{|FP|}{|FP| + |TN|} \quad (3.14)$$

$$F1 = \frac{2\,.\,precision\,.recall|}{|TP| + |TF|} \quad (3.15)$$

$$MCC(TP,TN,FP,FN) = \frac{(|TP|\,.|TN|) - (|FP|\,.|FN|)}{\sqrt{(|TP|+|FP|)\,.(|TP|+|FN|)\,.(|TN|+|FP|)\,.(|TN|+|FN|)}} \quad (3.16)$$

The performance assessment excluded seed nodes from consideration. Nevertheless, the performance assessment included all edges in the reconstructed pathways since we did not utilize interactions as an input. For a given interactome G(V, E) and a seed node-set (V_I), we reconstructed a pathway, R(V_R,E_R) to estimate a ground truth pathway T(V_T,E_T) where V_T, V_R, and $V_I \subseteq V$, and E_T and $E_R \subseteq E$. Node-level true positives (TP_V) and edge-level true positives (TP_E) are derived from $|V_R \cap V_T|$ and

$|E_R \cap E_T|$, respectively. $|V\backslash(V_R \cup V_T)|$ and $|E\backslash(E_R \cup E_T)|$ provide node-level true negatives (TN_V) and edge-level true negatives (TN_E). False positives, FP_V and FP_E, are equal to |VR \VT| and |ER \ET| respectively, while false negatives, FN_V and FN_E, are |VT \VR| and |ET \ER|.

We executed principal component analysis (PCA) to determine the major scores for the highest variation across all paths[196]. The whole set of performance data, including both edge- and node-based scores, was statistically evaluated by independently grouping the reference interactomes and reconstruction methods.

3.1.4.2. Parameter tuning

Reconstruction methods were optimized separately for each reference interactome. The selection of parameters was based on the use of the Wnt, TCR, TNFα, and TGFβ pathways available on NetPath. The nodes belonging to each pathway were randomized and divided into five-fold separately. We removed each fold from the complete pathway node list and then executed network reconstruction methods with the remaining folds. The parameters of the reconstruction methods were individually adjusted for each reference interactome to optimize the F1 score, as described in **Formula 3.15**. In the context of the all-pairs shortest path (APSP) algorithm, the determined shortest pathways among initial node sets were directly included in a reconstructed pathway without any parameter tuning. Considering each reference interactome, the parameters were adjusted within the specified range, as shown in **Table 2**, for the PRF, HDF, and PCSF algorithms. PRF and HDF parameter sets were changed in a two-dimensional grid by averaging the parameter settings for the 10 highest F1 scores. A union of parameter sets achieving the highest coverage of the seed nodes, V_I, for each pathway was the basis for determining the PCSF's optimal parameter sets.

Table 2: Tunning ranges of parameter sets in PageRank flux (PRF), heat diffusion flux (HDF), and prize-collecting Steiner Forest.

Reconstruction algorithm	Parameter	Range	Increment
PRF	Damping factor (λ)	0–1	0.05
	Flux threshold (τ)	0–1	0.05
HDF	Heat diffusion rate(α)	0–1	0.05
	Flux threshold (τ)	0–1	0.05
PCSF	Dummy edge weight (ω)	0–5	0.5
	Edge reliability (β)	0–5	0.5
	Degree penalty (γ)	0–10	0.5

3.2. Results

3.2.1. Systematic evaluation of reference human interactomes

The quality and coverage of the reference interactome heavily influence the effectiveness and accuracy of network reconstruction techniques. Consequently, a comprehensive investigation displayed the characteristics of iRefWeb, PathwayCommons, HIPPIE, ConsensusPathDB, OmniPath, and STRING databases. Several databases, like iRefWeb, HIPPIE, ConsensusPathDB, and STRING, provide scores to quantify confidence levels in interactions. Initially, a comparison was conducted between the pairs of interactomes to assess their similarity concerning their node and edge sets. PathwayCommons has the largest network size, resulting in a comparatively higher proportion of node and edge overlap with other interactomes. iRefWeb, PathwayCommons, HIPPIE, and ConsensusPathDB exhibit the highest similarity across interactomes, as determined by the overlaps in their nodes and edges (**Figure 6A**). In contrast, STRING and OmniPath interactomes have fewer numbers of shared nodes and edges. It is essential to acknowledge that we selected only experimentally known interactions of STRING that comprises more than one million interactions in the home ground.

For the comparative analysis, only experimental interactions were considered, resulting in an interactome of comparatively smaller size with edges of medium or high confidence. Before using network reconstruction methods, acquiring more reliable interactions in a reference interactome is fundamental, based on confidence scores or experimental methods. This step is essential in mitigating the influence of false positives. Network reconstruction methods mainly leverage the edge confidence scores and the topology of the reference interactomes during the propagation or

optimization, affecting the accuracy of the resulting network. **Figure 6B** displays the number of edges in each reference interactome by classifying edges into three categories: low, medium, and high confidence, determined by the interaction scores. Most interactions in ConsensusPathDB are characterized by increased confidence, while ones in HIPPIE and iRefWeb are distributed in medium and low confidence ranges. HIPPIE and iRefWeb use the MINT-inspired (MI) confidence score computation[197], while ConsensusPathDB employs the IntScore tool[49]. We recomputed the confidence scores in STRING based only on the experiment and database scores[198]. Both PathwayCommons and OmniPath lack the provision of confidence scores. The use of diverse scoring strategies results in variations in the distribution of confidence scores across the interactomes. The correlation coefficient between scores obtained from the HIPPIE and iRefWeb databases is very high ($r = 0.67$, $p < 0.05$). Conversely, the correlation between confidence ratings in iRefWeb and ConsensusPathDB is considerably low ($r = 0.25$, $p < 0.01$). This discrepancy may result from the distinct scoring method. (**Figure 6C**). The MI-Score metric takes into account homologous interactions, the technique of detection, and the number of publications about the interactions. On the other hand, IntScore combines topological features, evidence from the literature, and similarity in protein annotation.

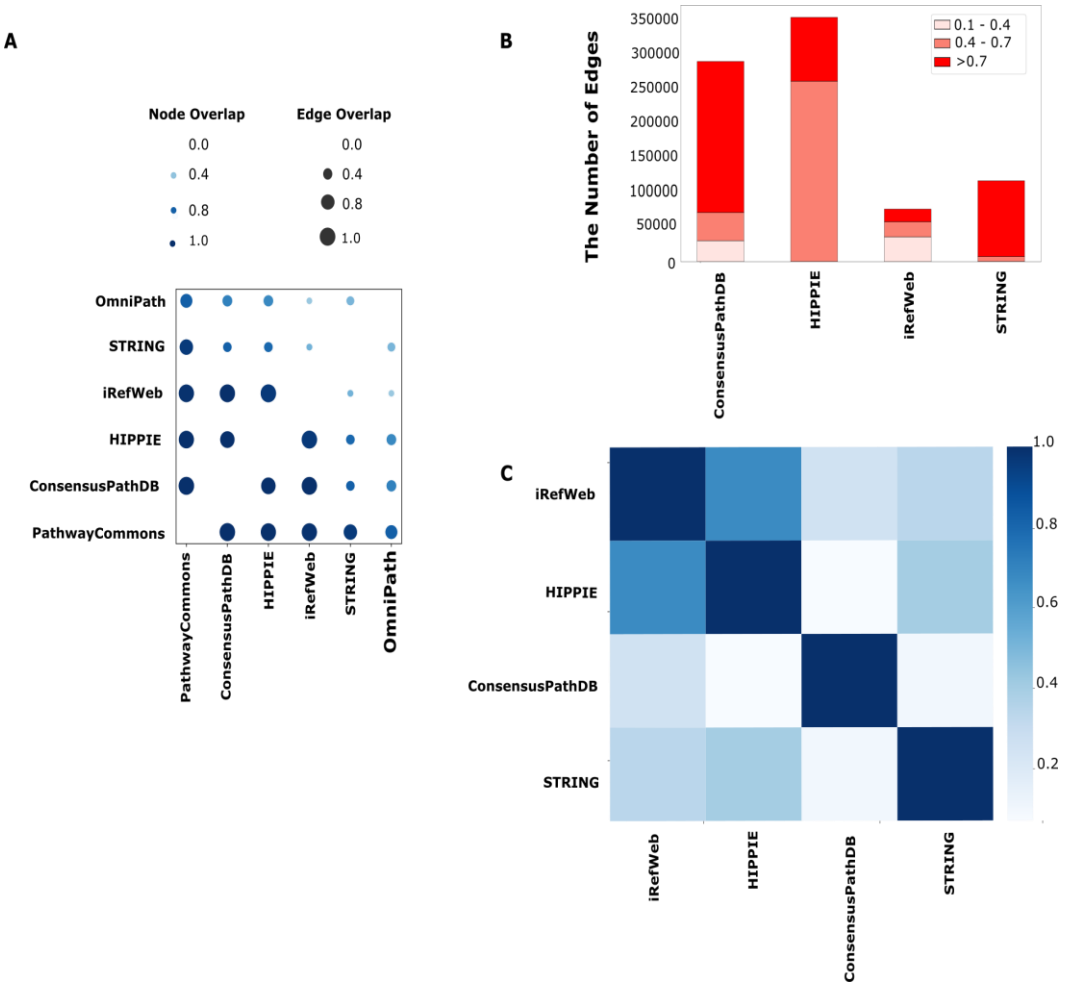

Figure 6: A comparative analysis of the reference interactomes **A)** Commonalities at the node and edge levels between given interactomes: A light-to-dark blue color scale displays the node overlap score, while the circle size depicts the edge similarity scores; the more significant the circle, the more prominent the similarity. **B)** The Confidence scores for each interactome are classified into three categories based on their confidence levels: low confidence (ranging from 0.1 to 0.4), medium confidence (ranging from 0.4 to 0.7), and high confidence (ranging from 0.7 to 1.0). PathwayCommons and OmniPath are not demonstrated here due to the lack of a confidence score for their edges. Edges with low confidence ratings are mostly seen inside the iRefWeb database, with only some segments of edges within the ConsensusPathDB database displaying low confidence scores. Conversely, the filtered STRING and HIPPIE databases do not include any instances of edges with low confidence scores. **C)** Various approaches for calculating confidence scores are used in interactomes As a result of evaluating their shared edges, the heatmap visually represents the correlation coefficients between confidence ratings across interactomes. The darkest blue represents the highest correlation between the HIPPIE and iRefWeb databases. Both interactomes use the same approach, MI scoring, for calculating confidence scores.

Despite helping filter out false positives, confidence scores alone are insufficient to mitigate bias within interactions. Consequently, we conducted further analysis on the interactomes, considering the inclination towards extensively researched proteins and using several indicators such as the number of publications about the proteins, including cancer driver genes, and the availability of structural information for the interactions. Well-studied proteins like TP53 and EGFR in the interactomes have hundreds of high-confidence interactions[199,200]. Indeed, a trade-off exists between the confidence scores of specific proteins and systematic research bias. We examined the number of publications and the degree centrality of proteins within each reference interactome to investigate the potential correlation between protein centrality and research attention using log-based values to find out their correlation (**Figure 7A**). The strongest correlation in PathwayCommons implies a bias toward well-studied proteins among these interactomes. iRefWeb, STRING, and OmniPath have a moderate correlation between the degree and the number of publications, which implies relatively less biased interactions.

The use of network reconstruction methods includes the identification of disease-associated pathways, particularly in cancer, by predicting markers, associated genes/proteins, and action mechanism of drugs and clustering or specifying patients clusters[19,110,201,202]. Therefore, we explored the cancer-driver genes (CDGs) in each interactome. CDGs provide a growth advantage to the tumor cells and induce alterations in signaling cascades and cell pathways. Identifying CDGs plays a crucial role in categorizing, characterizing, and advancing therapeutic interventions for tumors[203–205]. **Figure 7B** illustrates that CDGs have considerably more publications than the other proteomes ($p<0.01$). The identification and characterization of driver genes and their associated interactions play a crucial role in the precise reconstruction of driver pathways in cancer. The degree of CDGs is the introductory knowledge about their function in cancer progression since interactions of proteins determine their affecting mechanisms. CDGs in PathwayCommons have a higher number of connections compared to other interactomes (**Figure 7C**).

The structural identification of PPIs can be the most reliable and precise source and informative, uncovering the binding sites, domain contacts, and many more[206–208].

However, experimentally known PPIs are the only drawback in datasets. The number of protein complexes can only account for around 16% of the whole interactome, despite the exponential growth in PDB enabled by X-ray, CryoEM, and NMR methods[181,209,210]. We further examined each interactome in light of the representation of structurally annotated interactions. We used the PDB and Interactome3D complexes for this objective. Our research **(Figure 8A)** shows that HIPPIE has the most outstanding coverage of structurally known protein-protein interactions. PathwayCommons and ConsensusPathDB come after HIPPIE. The lowest coverages are in iRefWeb, OmniPath, and the filtered STRING interactome.

Generated subnetworks should be biologically meaningful so that their downstream analysis can identify proper biological functions, signaling cascades, and pathways[211,212]. Therefore, we investigated the coverage of interactomes by using curated pathways extracted from KEGG, one of the most widely used databases for pathway annotations. We discovered that PathwayCommons covers KEGG pathways much more and filtered STRING than ones by iRefWeb (**Figure 8B**).

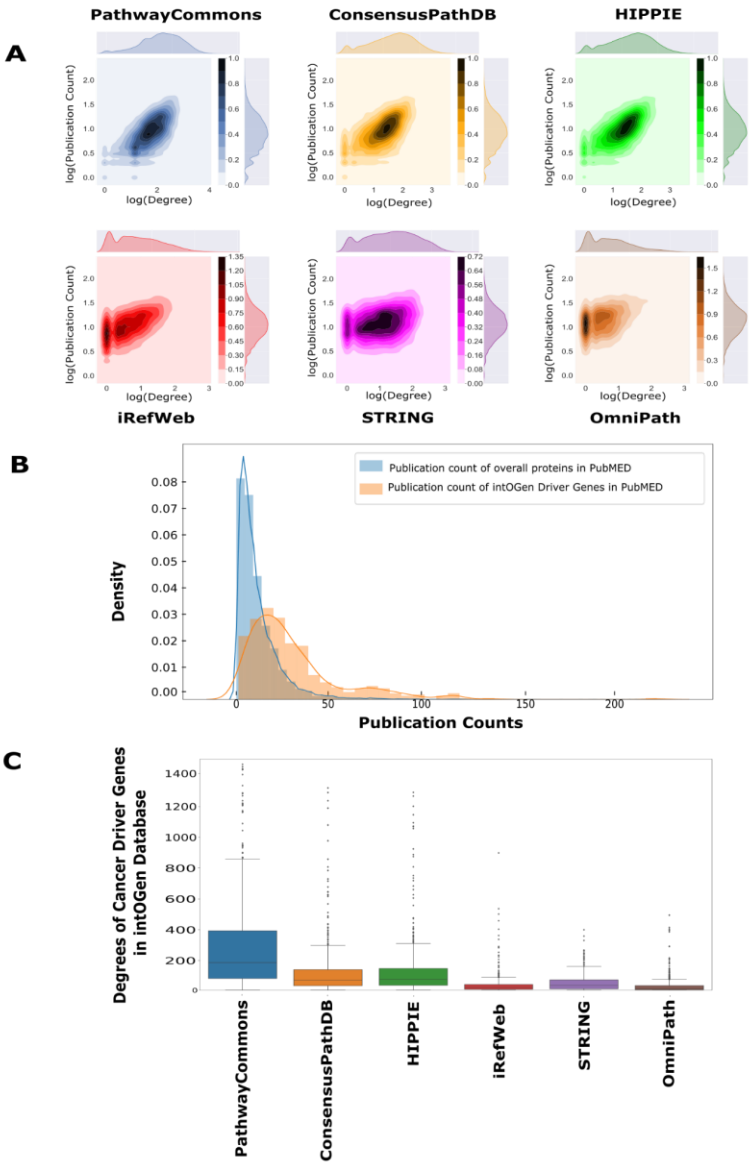

Figure 7: Correlation between publication counts and degrees in the context of interactomes **A)** Graphs depict the distribution of publications and degrees for each interactome on a log-log scale, exhibiting a power-law pattern. All interactomes have a positive association between degree and publication number. Notably, PathwayCommons, HIPPIE, ConsensusPath, and iREF have hubs extensively researched in the literature. In contrast, the hubs identified in iRefWeb and OmniPath exhibit a lack of well-studied proteins, as shown by their respective p-values (<0.001) and correlation coefficients with other databases ($r_{PathwayCommons}$ = 0.622, $r_{ConsensusPathDB}$ = 0.556, r_{HIPPIE} = 0.614, $r_{iRefWeb}$ = 0.508, r_{STRING} = 0.250, and $r_{OmniPath}$ = 0.400). **B)** The distributions of the number of publications and cancer driver genes in the intOGen database are shown in blue and orange, respectively. The probability of cancer-driver genes (CDGs) is greater than the probability of well-studied proteins. **C)** The boxplot illustrates the distribution of driver gene degrees among interactomes; CDGs in PathwayCommons have more connections than other interactomes have. OmniPath and iRefWeb have a comparatively lower abundance of interactions belonging to CDGs compared to ConsensusPath, HIPPIE, STRING, and PathwayCommons.

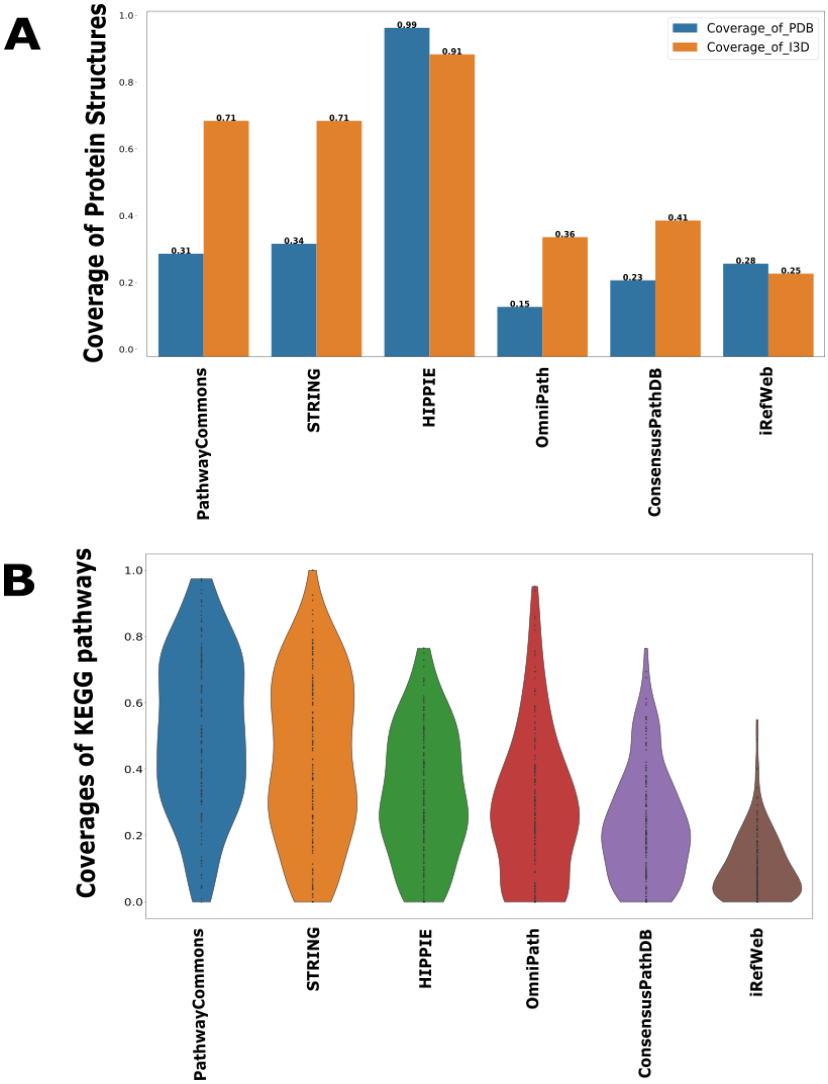

Figure 8: Coverages of known structurally known and curated interactions. **A)** Structural information is demonstrated in two groups: known interactions in PDB in blue and predicted interactions in Interactome3D in orange. **B)** The violin plot illustrates the distribution of overlaps between the interactions in KEGG pathways and each interactome.

3.2.2. Performance of network reconstruction algorithms

Each interactome has particular strengths and shortcomings, detailed under the title of the Systematic Evaluation of Reference Human Interactomes. To understand their performance, we've employed each interactome for each network reconstruction algorithm to assess the variation in their individual performance. The performance of four well-known network reconstruction algorithms, namely the all-pair shortest paths (APSP), personalized PageRank with flux (PRF), heat diffusion with flux (HDF), and

the prize-collecting Steiner Forest (PCSF) algorithms, were evaluated with the benchmark dataset of 32 curated pathways obtained from NetPath. Four pathways are used for parameter adjustment, while the remaining 28 are used for performance assessment.

We collected performance measurements at both the node and edge levels for each combination of interactomes and reconstruction techniques across all pathways. The performance of node-level analysis is more resilient to variations in interactomes, or pathways compared to edge-level analysis in each method. The most significant variation is in the F1 scores at the edge level, where the trade-off between recall and accuracy values exhibits considerable variability among routes and interactomes (**Figure 9**). The F1 scores ($p < 0.001$) and precision ($p < 0.001$) scores of the reconstructed pathways using PathwayCommons exhibit mainly lower values compared to the scores obtained from the other interactome (**Figure 10A**). The edge-level MCC, used for binary classification tasks using unbalanced data, has the second largest variation[213,214]. The outcome suggests that the algorithms exhibit suboptimal performance when applied to a reference interactome of considerable size, primarily due to the prevalence of false positive interactions overshadowing the actual positive interactions. Based on the F1 scores and precision values, our analysis did not provide any statistically significant differences in performance when using the HIPPIE, ConsensusPathDB, OmniPath, or iRefWeb interactomes. Consequently, we used HIPPIE as a reference interactome for further evaluations due to its well-balanced characteristics, including coverage of structurally known interactions, as determined in the reference network comparison.

The analysis of edge-based performance ratings revealed that APSP exhibits considerably lower accuracy values ($p <0.001$) and higher recall values than other reconstruction techniques when evaluating performance across all paths. There is a lack of statistically significant variation in precision values observed across HDF, PRF, and PCSF, as seen in **Figure 10B**. There is no significant difference in the recall values of the reconstructed pathways between HDF and PRF. However, PCSF exhibits considerably greater recall scores compared to HDF and PRF ($p < 0.001$) (**Figure 10C**). Including all shortest routes between the seed nodes in APSP causes a decrease

in accuracy values and an increase in recall values. The boosted FPR seen in APSP suggests that the number of false positive edges surpasses the number of accurate positive edges. Hence, the F1 scores of the APSP-reconstructed pathways exhibit a statistically significant decrease compared to those of other methods ($p < 0.001$) (**Figure 10D**). In contrast, pathways reconstructed by PCSF have reasonably high recall and precision scores and the greatest F1 score, optimizing the precision and recall values. Notably, the range of recall scores seen in the reconstructed pathways with the PCSF approach is not as diverse as in other methods because of the intersection of multiple solutions derived from various parameters. The Omics Integrator 2, using the PCSF algorithm, assembles an ideal forest as its primary output. Additionally, it generates an augmented forest that encompasses all the edges that connect the nodes included in the optimal forest. The final network of PCSF was formed by intersecting the augmented forests generated from multiple parameters. Thus, incorporating an additional edge into the final network was executed with a high level of strictness. We computed the Jaccard similarity matrix among HDF, PRF, and PCSF to demonstrate the variation in the edge-level performance across the reconstructed pathways (**Figure 10E**)[215]. The PCSF algorithm penalizes the nodes with strong connectivity, diminishing the influence of well-studied or hub nodes in the reconstructed networks.

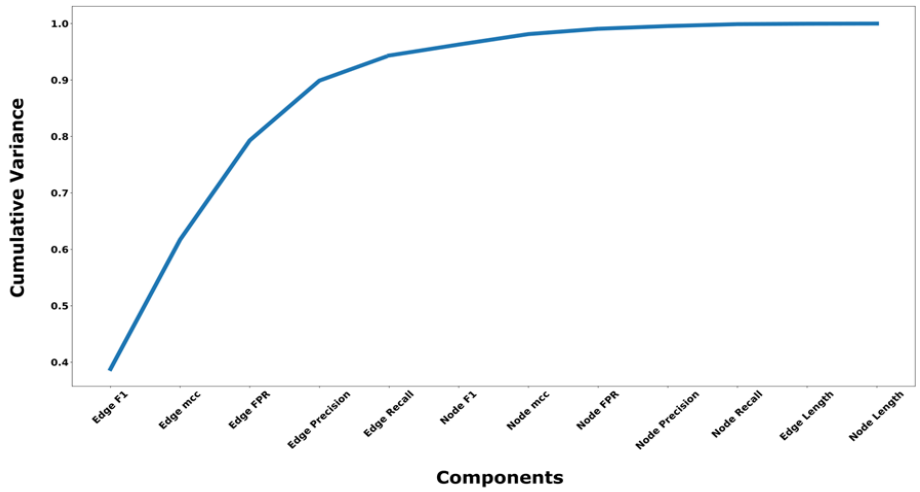

Figure 9: Principal Component Analysis (PCA) over edge-based and node-based scores reveals that edge-based scores explain more than 90% of the variance.

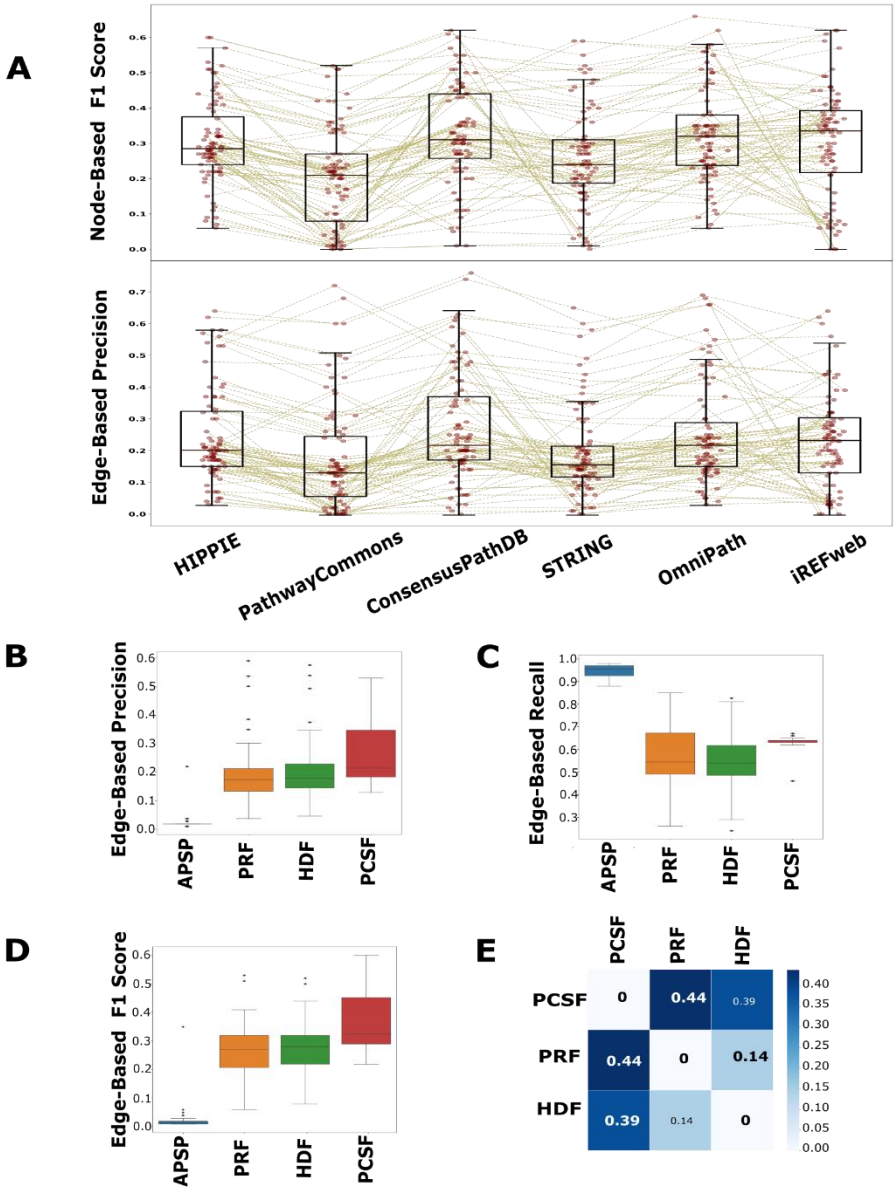

Figure 10: Performance assessment of each interactome and method in pathway reconstruction. **A)** The Boxplot of edge-based precision and F1 scores across different interactomes reveals that PathwayCommons and STRING exhibit considerably lower scores than the other interactomes. At the same time, there is not any distinct difference among HIPPIE, ConsensusPathDB, OmniPath, and iRefWeb. The performance score of each reconstructed network is illustrated with red points in the boxplots. Brown lines connect the performance scores of the same pathway across the interactomes. **B)** Edge-based precision, **C)** edge-based recall, and **D)** edge-based F1 scores are demonstrated for an individual reconstruction algorithm. **E)** The reconstructed pathways were assessed considering HDF, PRF, and PCSF. The heatmap displays that the reconstructed pathways by PCSF are different, having %44 and %39 different edges, respectively, than the ones reconstructed by PRF and HDF.

3.2.3. Reconstruction of the notch pathway

The first case study, the Notch signaling pathway, is crucial to cell fate determination, regulating differentiation, apoptosis, proliferation, and morphogenesis. Cancer studies cover this signaling cascade and its crosstalking pathways[216–218]. We did not consider the APSP method due to many false positives, while PRF, HDF, and PCSF results for the Notch pathway are demonstrated in **Figure 11**. Notch receptors, single-pass transmembrane proteins, initiate a signaling cascade by receiving signals from transmembrane ligands such as JAG1, JAG2, DLL1, and DLL4. Our seed nodes covered Notch receptors and CNTN1, JAG2, and DLL4. During propagation, each reconstruction algorithm determined JAG1 and interaction between Notch receptors and ligands apart from DLL. PCSF proved superior performance in recovering low-degree nodes, such as CNTN1, WDR12, LEF1, RBX1, SIN3A, and other true positives. However, it could not include several additional nodes, such as AKT1, SKP1, SPEN, and TCF3. Furin–Notch receptors, successfully identified by PCSF, regulate the Notch pathway in cancer progression[219]. HDF and PRF mostly identified the interaction between highly connected nodes, such as MAML1 and Notch receptors, while fainting to construct interactions between nodes with low degrees, such as JAK2 and WDR12.

The Notch signaling pathway interacts with other critical pathways in cancer, such as PI3K-AKT-mTOR and JAK-STAT signaling pathways. These interactions, providing crosstalk, are mainly mediated by the nodes with low-degree and high betweenness centrality in reference networks[220]. PCSF identified intermediate or hidden nodes mediating crosstalk while being insufficient to add their interactions. In the reconstructed Notch signaling pathway, we correctly identified the PIK3R1-Notch1-LCK interactions, but we could not find the PIK3R2-AKT interaction. Similarly, in the JAK-STAT and Notch pathway crosstalk[221,222], we accurately found intermediate nodes such as JAK2, HES1, and HES5, but we failed in recovering their interactions with STAT3 in the PCSF-reconstructed pathway.

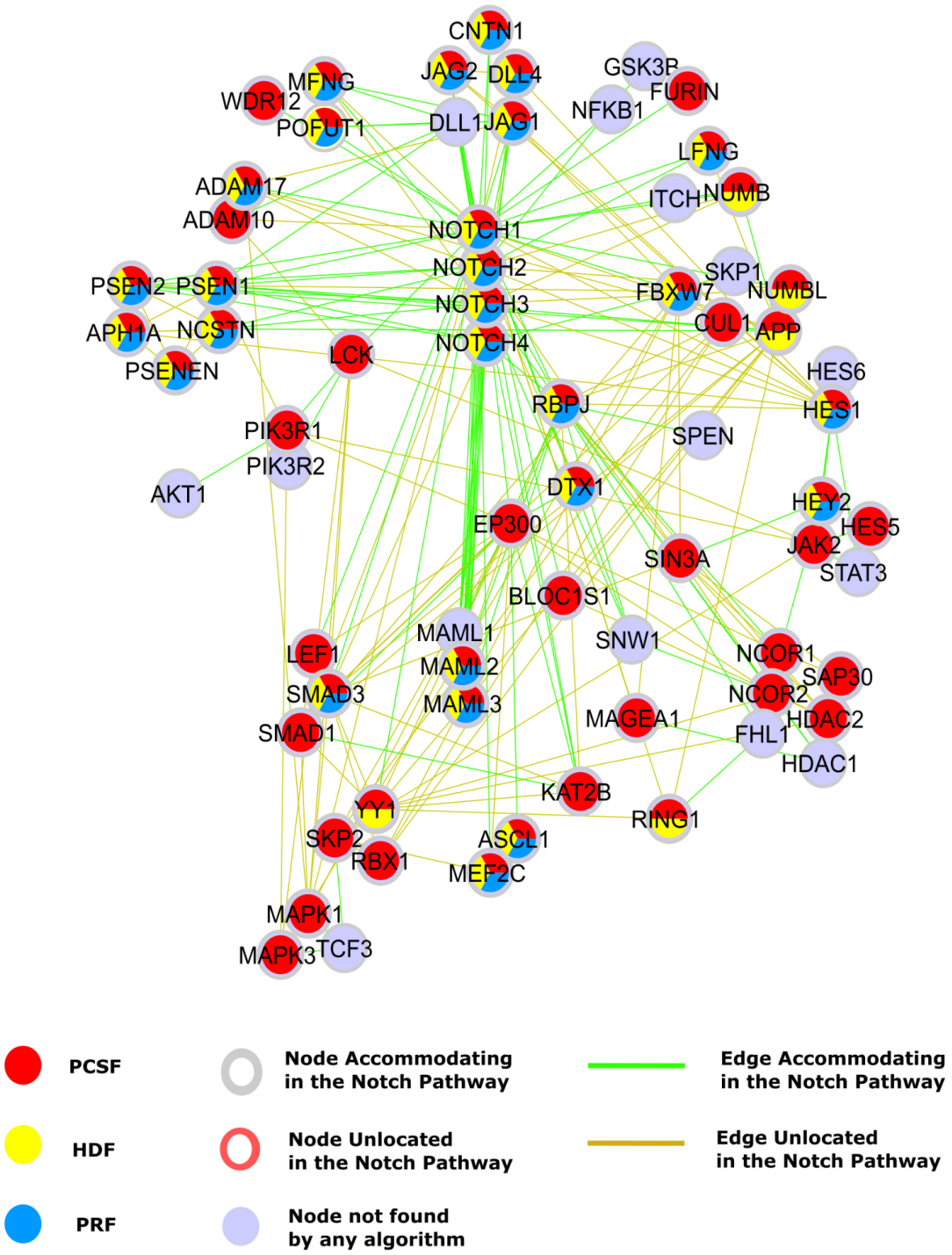

Figure 11: Reconstructed Notch pathway. Nodes that are present in the pathway but are not found by any algorithms are colored light blue. Nodes that PCSF, PRF, and HDF find are colored red, yellow, and cyan, respectively. Green edges are present in the Notch pathway in NetPath, while incorrectly included edges by any algorithm are shown in brown.

CHAPTER 4

pyPARAGON: COMBINING NETWORK PROPAGATION WITH GRAPHLETS TO INTEGRATE MULTI-OMICS DATA

Properly integrating and translating multi-omics datasets into interpretable information is challenging owing to data sparsity[223], missing data points[42,224], and computational complexity[29], as discussed in the literature review. Network-based algorithms enable addressing these issues and deciphering causal relationships between omics components[7,95,225]. These approaches ultimately acquire a network model that may depict changes in disease models or pharmacological treatments using topological and statistical characteristics. The merit of using global and local network characteristics (such as degree distribution and clustering coefficients) for propagation or inference is constrained when dealing with this kind of sparse data. Therefore, the frequencies of motifs (recurring subgraphs) might provide a more elucidating approach to unveiling complex cellular networks[167].

Graphlets, non-isomorphic small connected subgraphs, are found in higher proportions in the reference interactome and are linked to particular functions[162,163]. Another obstacle arises from highly interconnected and multifunctional proteins, namely hub proteins dominating the final network and obscure context-specific networks[226]. These proteins can potentially introduce non-specific interactions into the network models due to the small-world nature of reference interactomes. Therefore, the use of network motifs, graphlets, or module identification may enhance the context-specific elements of models[7,132,225].

This thesis proposes that utilizing network motifs instead of individual protein connections provides a more accurate representation of signaling networks and minimizes the incorporation of irrelevant interactions. We have proved that graphlets effectively decrease the complexity of reference interactomes by eliminating non-

specific and highly interconnected proteins and their interactions. pyPARAGON (PAgeRAnk-flux on Graphlet-guided network for multi-Omics data integratioN) outperformed chosen state-of-the-art methods, such as Omics Integrator 2[34] and PathLinker[137], regarding node propagation and edge inference on the benchmark set of cancer signaling pathways.

4.1. Methods

4.1.1. Overview of pyPARAGON as a hybrid network inference framework

Combining more than one approach can be more successful in multi-omics integration than depending on a single method alone[169,227]. pyPARAGON is an innovative method integrating graphlets with network propagation using the personalized PageRank algorithm. It then selects interactions based on edge flux calculation to tackle the issues in network modeling effectively. pyPARAGON runs in three steps (**Figure 12A**): i) the construction of Graphlet-guided network (GGN), ii) propagation and edge scoring, and iii) selecting highly scored edges on GGN.

Cutting-edge techniques consider an immediate connection between two nodes in the reference network and features of nodes (such as degree, betweenness, proximity, and eigenvector centralities). The GGN creation stage of pyPARAGON utilizes an unsupervised strategy to find a core area in the reference interactome by merging considerable frequent graphlets with 2-4 nodes (**Figure 12B**). it is expected to find few direct interactions between the genes/proteins of interest. Intermediate nodes are necessary to connect them and construct a coherent network structure. Therefore, we restricted graphlets with more than two nodes that may have an intermediary node. Intermediate nodes are defined as the nodes with the most connections to the seed nodes in a graphlet.

Aside from constructing GGNs, the personalized PageRank algorithm also spreads signals from seed nodes across the reference interactome. The node weights and their degrees, and edge confidence ratings are integrated into a unified function to compute edge fluxes[173]. Within this function, the degree component punishes proteins with a high level of connectivity that can be arbitrarily inserted in a final subnetwork. Ultimately, we establish connections between edges with significant flux scores in

GGN (**Figure 12C**). For interpretation, using the Louvain community detection method[151], pyPARAGON also reveals communities operating in distinct biological processes or pathways (**Figure 12D**). Subsequently, the hypergeometric test in pyPARAGON identifies context-specific annotations[228]. In this way, we reveal not only hidden connections between initial nodes but also significant context-specific pathways.

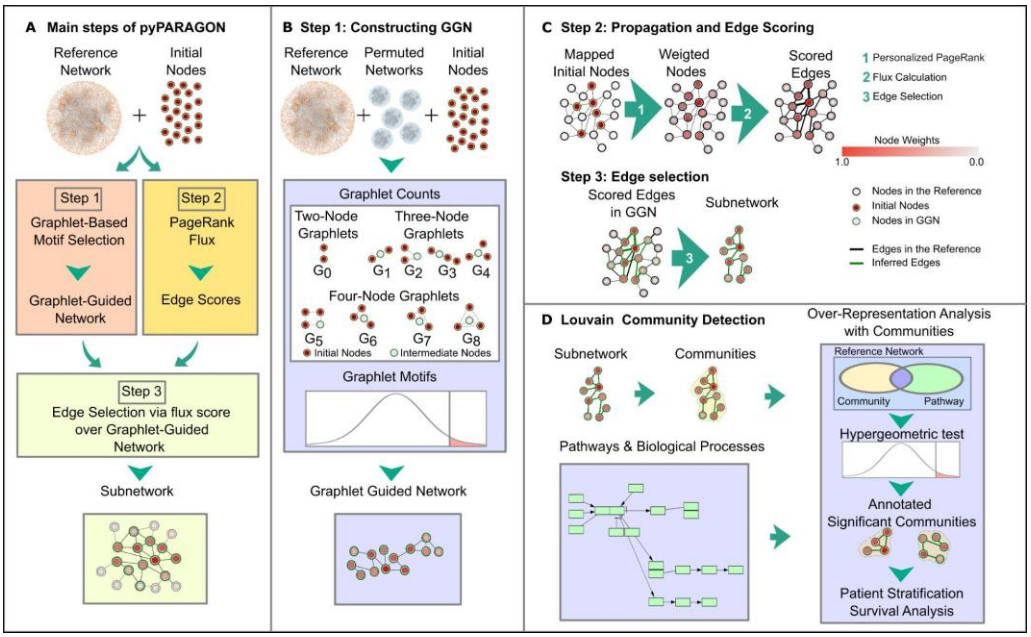

Figure 12: The overview of pyPARAGON **A)** pyPARAGON runs in three steps: i. GGN construction (light red boxes); ii. Propagation and edge scoring (yellow boxes); iii. Selecting highly scored edges in GGN (green boxes). **B)** We investigated nine non-isomorphic graphlets (G_0-G_8) composed of 2, 3, and 4 nodes for GGN. Except for G_0, each graphlet covers at least two seed nodes (red circles) and one intermediate node (white circles) that connects the seeds in the center of the orbit. **C)** By random walking from weighted initial nodes in the reference network, the Personalized PageRank algorithm assigns a weight to each node. Computed edge fluxes were used as the edge scores in the reference interactome. High-scoring edges in GGN formed the final subnetwork. **D)** pyPARAGON employs the Louvain community detection method, based on network topology, to divide the inferred network into functional units. Significant biological processes and pathways in each community were found by a hypergeometric test.

4.1.2. Network inference tools

4.1.2.1. pyPARAGON

During the construction of GGN, we parsed 2-, 3-, and 4-node-graphlets (G_0, G_1, G_2, …, G_8, as seen in **Figure 12B**), which are small non-isomorphic subgraphs. An

isomorphism of graphlets between two subgraphs, X(V_X, E_X) and Y(V_Y, E_Y), is established with bijections, a one-to-one equality, between V_X and V_Y[162]. While searching graphlets, we looked through the graphlets to find an intermediate node in one of the orbits with the highest degree and seed nodes in the remaining orbits. The reference network is defined as R(V_R, E_R, c(e)), where V_R, E_R, and c(e) are nodes, undirected edges, and the confidence score of an edge, respectively. Furthermore, we computed the frequencies of graphlets in 100 permuted networks using the same seed node set[169,170]. We implemented the z-test to compare the frequencies of targeted graphlets in the reference and permuted networks ($p<0.05$, z-score>1.65). The graphlet-guided network (GGN), denoted as G(V_G, E_G), where $G \subseteq R$, is formed by the union of graphlet motifs, which represents notably frequent graphlets.

In the propagation and edge scoring steps, we implemented the personalized PageRank (PPR) algorithm and flux calculation detailed in the previous chapter, network reconstruction part. The final edge score (f(e)) is defined with **Formula 3.8**. After mapping edge scores into GGN and applying **Formula 3.9**, we select the most weighted edges to infer the context-specific network C(V_C, E_C), where $E_C \subseteq (E_G, >)$ and $V_C \subseteq V_G$.

We use the Louvain method to detect communities or modules in context-specific networks, a fast and heuristic method composed of two iterative steps. (1) Assigning each node to its community; (2) Interchanging neighbor nodes to find the maximum modularity until no positive gain is achieved[151].

We parse the functional communities using the hypergeometric distribution. The probabilities of communities are examined with the prior knowledge using **Formula 4.1** as follows:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{M-N}{n-i}}{\binom{N}{n}} \qquad (4.1)$$

where M is defined as the population size, the number of genes in the given reference network; N is the number of genes/proteins in the prior knowledge; n is the number

of genes/proteins in the community; and k is the number of successfully identified genes.

4.1.3. Interactomes and datasets

We used different interactomes as references: HIPPIE v2.2 (15,861 nodes, 345,770 edges), HIPPIE v2.3 (19,437 nodes, 774,449 edges)[8], and ConsensusPathDB v35 (18,178 nodes, 516,211 edges)[175]. 18 cancer signaling pathways with more than 50 proteins were retrieved from NetPath as a benchmark[195]. We established the seed node set for 8 cancer types, namely bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), esophageal carcinoma (ESCA), head and neck squamous cell carcinoma (HNSC), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), and uterine corpus endometrial carcinoma (UCEC). Out of the 1,289,655 mutations in 3,759 individuals, we identified the 300 genes with the highest mutation occurrence. The mutant dataset encompasses a range of cancer genomics initiatives, including TCGA and GENIE[229]. We obtained ground-truth node sets for cancer types from IntOGen, composed of the 3333 driver mutations on 568 genes[182].

4.1.4. Performance assessment of network inference tools

We assessed pyPARAGON by performing pathway reconstruction in NetPath, inferring specific cancer networks, and comparing its performance with Omics Integrator 2 and PathLinker 1.4.3. The network inference tool requires a list of seed genes/proteins (initial nodes) as input, which should be tailored to the relevant biological context. Seeds may be acquired by several methods, including but not limited to omics research, pharmacological perturbation analysis, or disease-associated proteins. Firstly, we reconstructed cancer signaling pathways in NetPath and disease network models of 8 cancer types. Each pathway in NetPath was shuffled separately, dividing their nodes into two equal portions, repeated five times. One portion was utilized for seed node sets, while the remaining portion was attempted to be in a final reconstructed pathway by covering their edges. During modeling cancer types, the driver genes were randomly partitioned into five equal segments for each type of cancer. Each segment was extracted from the genes with the highest frequency of mutations. Next, we used the remaining frequently mutated genes as initial nodes

to model each cancer type. We attempted to predict extracted driver genes using different inference tools during assessments.

4.1.4.1. Network inference tools

Omics Integrator 2 utilizes the prize-collecting Steiner Forest algorithm integrating the objective function (**Formula 3.11**), detailed in the previous chapter.

PathLinker computes the k-highest scored short pathways between seed nodes without a loop in the reference network. The path score, denoted as *W*, is calculated by multiplying the edge weights along the path[137]. The cost of a path is calculated with **Formula 4.2**.

$$C_{uv} = \{-log(W_uv)\ if\ u, v \in V \setminus \{s, t\}, 0\ if\ u\ =\ s\ or\ v = t\ \} \qquad (4.2)$$

where *s* and *t* are, respectively, a source and a target for each node, x ∈ S. The cost of a path is the sum of the costs of the edges in the path.

4.1.4.2. Assessment metrics

We performed a comparative analysis of GGN and reference networks based on topological features and network metrics, such as node and edge counts, highly connected nodes with more than 200 interactions, average node degrees, and diameters.

We computed precision, recall, F1 scores, and area under the precision-recall curve (AUPRC) for each pathway and cancer-specific network[38]. We generated a series of parameter sets for tools using grid search. In Omics Integrator 2, the parameter sets were defined as follows: the dummy edge weight (ω) was varied, the edge dependability (β) ranged from 0 to 5 with increments of 0.5, and the degree penalty (γ) ranged from 0 to 10 with increments of 1. Similarly, we assessed the efficiency of PL by varying the value of k, which represents the number of shortest pathways, from 50 to 1000 in increments of 50. In contrast, we evaluated the performance of pyPARAGON by adjusting the damping factor (λ) and flux threshold (τ) within the range of 0.05 to 1, with increments of 0.05.

We measured the changes in the connectivity of proteins that are strongly interconnected between the provided reference network and GGN to assess the performance of GGN. In our study, we identified a set of proteins, denoted as H_R, that have more than 200 interactions in a reference network $(h_1, h_2, ..., h_n) \in H_R$ for a reference network, the highly connected proteins $(h_1, h_2, ..., h_m) \in H_G$ in GNN, and the highly connected proteins $(h_1, h_2, ..., h_p) \in H_P$ in the given pathway, $H_P \subseteq H_G \subseteq H_R \subseteq V_R$ The reduction ratio (RR) of the remaining highly connected proteins in GGN was separately calculated using **Formula 4.3**:

$$RR = \frac{log_{10} \sum_{i=1}^{m} \frac{deg_R(h_i)}{deg_G(h_i)}}{m} \quad (4.3)$$

deg(h) is the number of interactions of h, and m is the number of highly connected nodes in GGN. We computed RR of highly interconnected proteins for each signaling pathway.

4.1.5. Running time analysis of pyPARAGON

The running time analysis was conducted on a computer running Windows 11 and equipped with an Intel (R) CoreTM i7-10510U CPU, 1.80GHz, 2304 Mhz, 4 cores, and 16GB of DDR4 RAM. We separately measured running durations by altering two variables: the number of initial nodes and the size of the reference network. In the first case, we constructed a random geometric network composed of 15 000 nodes[230]. We used default parameter sets of pyPARAGON where graphlets (G_1, G_2, G_3, G_4, G_5, G_6, G_7, and G_8), $\lambda = 0.8$, and $\tau = 0.8$. The number of initial nodes was adjusted between 25 and 750, with 25 incremental. We randomly selected each initial node set 30 times. In the second case, we constructed random geometric networks composed of 1,000 to 20,000 nodes with an incremental increase of 1,000, repeated 30 times. During network inference, we appointed 100 random initial nodes from each own reference network and kept its running time.

4.2. Results

4.2.1. Network trimming via graphlets improves the reference networks

We recruited the NetPath dataset[195] as the ground dataset to build precise cancer signaling pathways and evaluate the effectiveness of pyPARAGON. Similarities in

topological features, predicted nodes, and edges in targeted networks and functional units are mainly assessed in the performance of methods[169,231]. After examining all graphlets in the reference interactomes, we determined that the most often occurring graphlets were G_2, G_5, G_6, G_7, and G_8 (**Figure 13A**). Direct interactions between input nodes are not significantly frequent in the given reference networks. On the other hand, these direct interactions get more significant in the presence of intermediate nodes interacting with G_0 and constructing G_2. Our observation shows that graphlets with at least one intermediary node connecting the seeds provide more precision than adding direct interactions between two seeds (i.e., G_0) in the GGN.

Each interactome has a distinct assessment and scoring system to integrate protein-protein interactions (PPIs) from various databases[38]. We measured different topological features in ConsensusPathDB, HIPPIE v2.2 and HIPPIE v2.3 (**Table 3**). GGN is a contextualized subset of the given reference interactome. The independent comparison of reference networks demonstrated that GGN, a trimmed interactome, significantly enhanced the similarities among reference networks (**Figure 14B**). The final GGN retains features of a scale-free network that follows a power law (**Figure 14C**). Although scale-free networks are still controversial in modeling approaches, various research found that functional biological networks are scale-free structure[232,233]. Scale-free networks are robust to the random loss of nodes, defined as error tolerance, and fragile to targeted worst-case attacks[234].

Table 3: Topological features of reference networks

	ConsensusPathDB		HIPPIE v2.2		HIPPIE v2.3	
	Reference Network	GGNs (Average)	Reference Network	GGNs (Average)	Reference Network	GGNs (Average)
The number of nodes	18 178	2240.22	15 861	1823.52	19 437	2805.78
The number of edges	516 211	8916.36	345 770	7137.52	774 449	11579
The number of highly connected nodes	1015	15.86	543	11.59	1812	21.11
Average node degree	56.8	7.26	43.6	7.27	79.6	7.52
Diameter	8	4.45	8	4.44	7	4.35

Crosstalk between cancer signaling pathways forepoints a variety of cellular functions such as cell survival, metastasis, or apoptosis[235,236]. Highly connected nodes, or hub proteins, with various functions in different pathways, provide crosstalks between pathways[237]. However, the specific functionality of these nodes in reference networks is another challenging issue. Thus, reducing the power of highly connected nodes is another benefit of GGN. Within HIPPIE version 2.3, we discovered 1812 highly connected nodes that had more than 200 interactions. The GGN construction in pyPARAGON effectively reduced the number of highly connected nodes while hosting associated ones (**Figure 14**). Moreover, highly connected nodes and their specific functionality are also critical false negatives. Reduction Ratios in GGN demonstrate that the remaining highly connected nodes in GGN mainly lost their interactions, while reduction ratios of highly connected nodes in true pathway sets remained low. Thus, GGN maintained the associated functional interactions in GGN while trimming irrelevant interactions in reference networks.

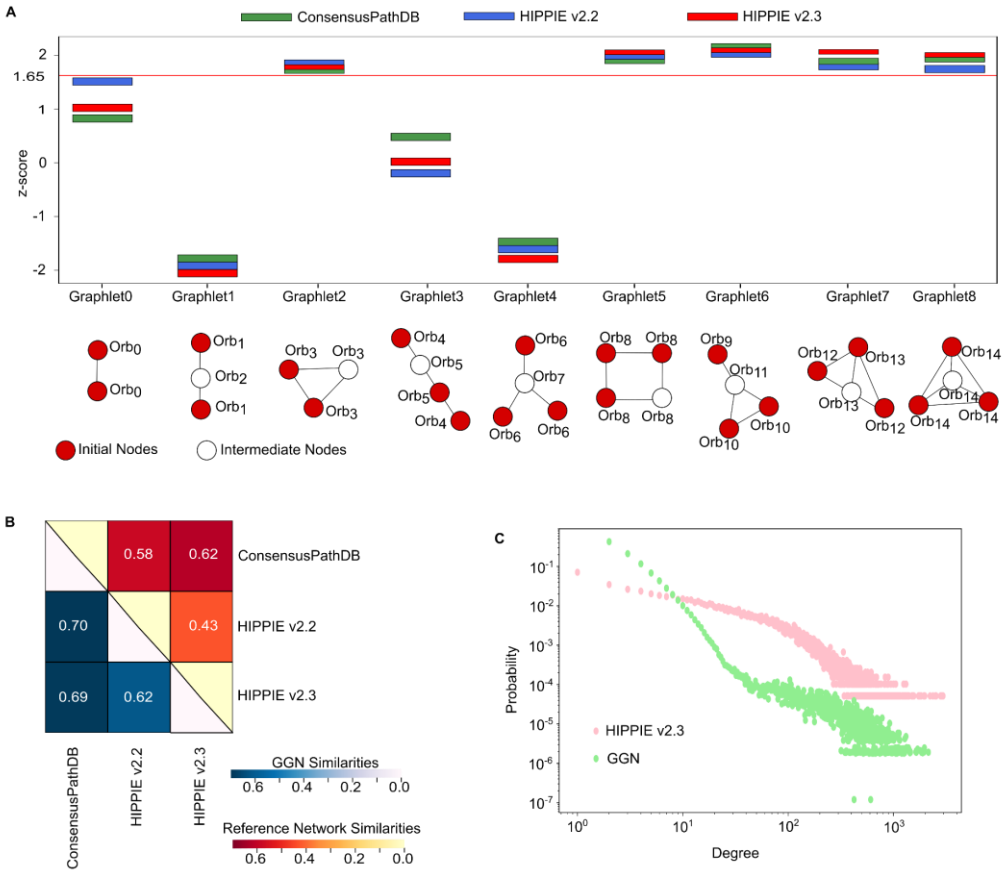

Figure 13: Graphlet-guided networks (GGN) optimize reference networks. **A)** Graphlets composed of 2, 3, and 4 nodes are constructed with initial nodes (red circles) coming from the given input and intermediate nodes (white circles). Intermediate nodes are the ones that have the highest connections to the seed nodes in the corresponding graphlet. We compared the frequencies of graphlets on different reference interactomes with their 100 permuted networks. Despite having different network sizes and properties, ConsensusPathDB (green), HIPPIE v2.2 (blue), and HIPPIE v2.3 (red) have similar graphlet motifs, such as Graphlets 2, 5, 6, 7, and 8 for signaling pathways in NetPath ($p<0.05$). **B)** The heatmap with the gradual color change highlighted the network similarities between reference networks (red) and GGNs (blue). The Jaccard Similarity Index was determined by dividing the number of common interactions by the number of merged interactions. The top-right section depicts network similarities, whereas the bottom-left section depicts average GGN similarities retrieved with the same initial nodes from NetPath. The construction of GGN results in more comparable and optimized networks. **C)** The pink and green distributions depict the degree probabilities of HIPPIE v2.3 and, GGN. A power law governs HIPPIE v2.3, a scale-free network. GGN was constructed with nodes at various degrees to avoid the noise of highly connected nodes by reducing their irrelevant interactions. GGN retains scale-free network features, as seen in biological networks.

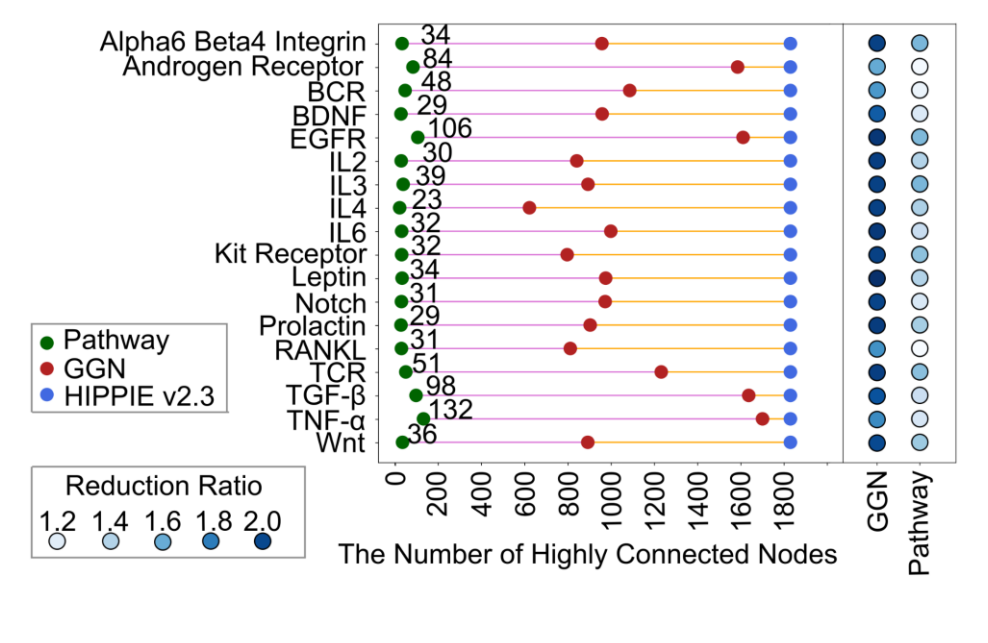

Figure 14: Graphlet-guided network trims reference interactome by removing some highly connected nodes and their non-specific interactions. Highly connected proteins are defined as the ones having more than 200 interactions in HIPPIE interactome (blue dots). The presence of these nodes in GNNs and NetPath pathways are shown for each signaling pathway (red and green dots, respectively). In the reference interactome, 1812 highly connected nodes are present. GGN selects a subset of these nodes that are highly specific to the pathways. The change in node degrees of remaining highly connected proteins in GGN was calculated as the reduction ratio and shown with a blue color scale. Highly connected nodes in the reference interactome that are present in pathways are included during reconstruction with a low reduction ratio in GGN, while the rest have a higher reduction ratio.

4.2.2. Performance of pyPARAGON on the reconstruction of cancer signaling pathways

Pathways are a particular course of serial actions among biomolecules in a cell, leading to a particular product (metabolomics) by triggering the assembly or disassociation of biomolecules, and a change in the cell by turning on or off genes. During the performance assessments of pyPARAGON, PathLinker, and Omics Integrator 2, we compared reconstructed pathways in node propagation and edge inference aspects. We evaluated the predicted nodes in the prior aspects, considering how to complete the missing piece of knowledge via propagation. In the latter, we look at how interactions are functionally true in edge inference. To assess performance, we used the area under the precision-recall (AUPRC) curve, which allowed us to evaluate how well each pathway's nodes and edges were recovered in the predicted networks. Our analysis

demonstrated that pyPARAGON had superior performance compared to PathLinker and Omics Integrator 2 in inferring signaling pathways throughout all pathways of NetPath, both at the node and edge levels (**Figure 15A**). As an additional metric, we recruited F1 scores, representing both precision and recall scores in balance. Indeed, there is a trade-off between precision and recall scores. While achieving a better recall score, we saw decreased accuracy in the inferred networks. Analysis of F1 scores showed that pyPARAGON and PathLinker have a higher efficiency in propagation, while pyPARAGON and Omics Integrator 2 outperform in network inference (**Figure 15B)**. On the one hand, highly connected reference networks weakened the propagation ability of Omics Integrator 2. On the other hand, the high number of interactions provided more robust interactions in Omics Integrator 2 that were resilient to hub nodes. Considering PathLinker, we saw that propagation of the seed nodes was more robust due to the use of multiple short paths but introduced many false positive interactions. As a result of biological networks being scale-free, many seed nodes have a propensity to be linked by hub nodes as shortcuts. For this reason, the random walk-based and shortest pathway techniques have the potential to result in the inclusion of false positive interactions[238,239]. However, applying penalties to highly connected nodes, such as using degree-dependent negative prizing in Omics Integrator 2 or penalizing nodes regarding the number of interactions, like calculating the PageRank flow to normalize the score in pyPARAGON, may decrease the number of false positive edges and enhance the F1-score in edge prediction.

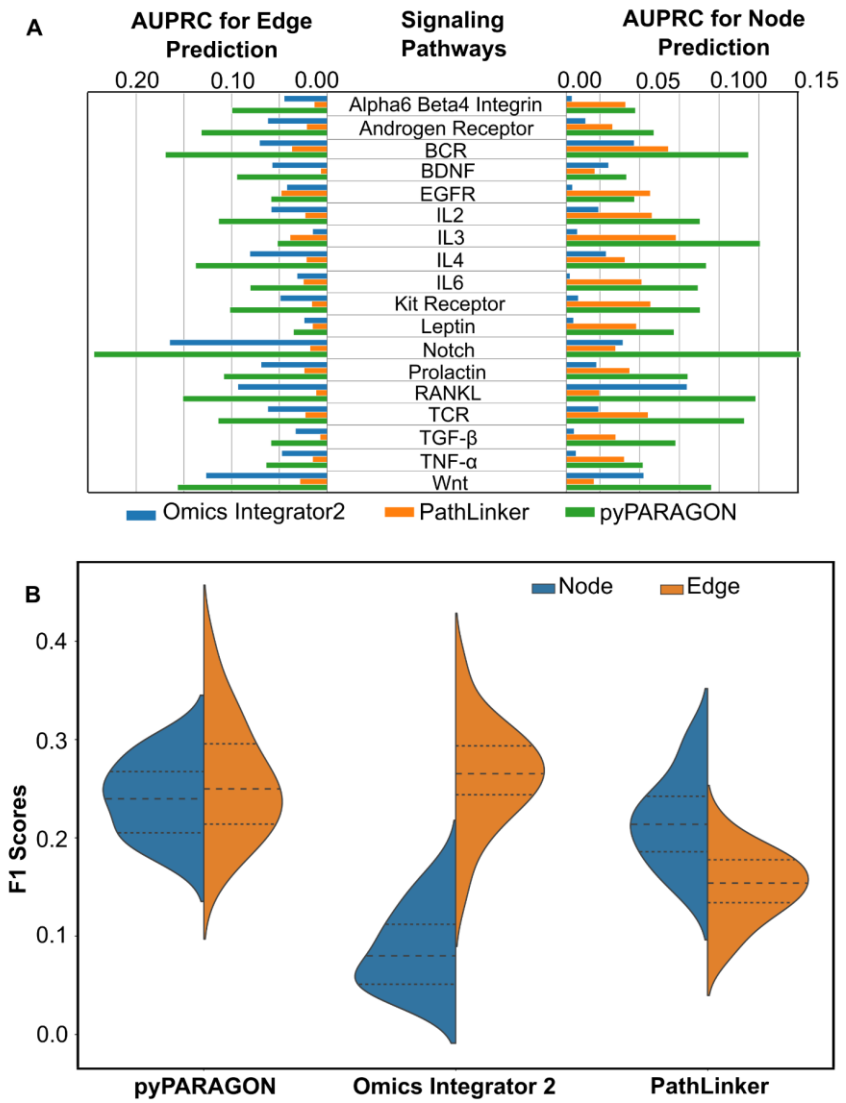

Figure 15: Outperformance of pyPARAGON over Omics Integrator2 and PathLinker**. A)** AUPRC of each tool (blue=Omics Integrator 2, orange=PathLinker, and green=pyPARAGON) in each pathway reconstruction is shown in bar-plot for the following tools: In all signaling pathways, pyPARAGON performed better than others in both node and edge predictions. **B)** Distribution of F1-scores for each tool across 18 pathways is shown for node (blue) and edge (orange) predictions.

4.2.3. Network-based modeling of cancer types

We constructed network models of 8 cancer types to assess the effectiveness of pyPARAGON with other selected tools. Initially, we appointed 300 most common mutations as seed nodes in eight cancer types. Known driver genes, retrieved from IntOGen, were utilized as an independent test set where we checked their presence in contextualized network models. Utilizing 5-fold cross-validation, we excluded the

shared proteins between the seed list and known drivers for each fold. Subsequently, we reconstructed cancer-type-specific networks using pyPARAGON, PathLinker, and Omics Integrator 2. Our findings demonstrated that contextualized networks, inferred by pyPARAGON, encompass more known driver genes than other tools in all cancer types. pyPARAGON achieved higher recall and precision scores (**Figure 16**). In the highly condensed reference networks, early termination of propagation between seed nodes is the cause of recovering fewer driver genes in cancer-type-specific networks inferred by Omics Integrator 2. Nodes with a high degree of connectivity provide network shortcuts rather than relying on signal cascades or motifs in these reference networks. In the PathLinker-contextualized networks, as a result of recruiting the multiple-shortest paths, intermediate nodes were mostly associated with highly connected nodes rather than particular driver genes. pyPARAGON employs the PageRank algorithm to propagate seed nodes beyond their surrounding reference interactome. Furthermore, by using graphlets, GGN creation eliminates potential "frequent flyers," allowing for more accurate prediction of driver genes. In general, pyPARAGON outperforms competing cancer driver networks in terms of precise prediction and may be fine-tuned for use in building tumor-or patient-specific networks and conducting network comparisons based on network similarity.

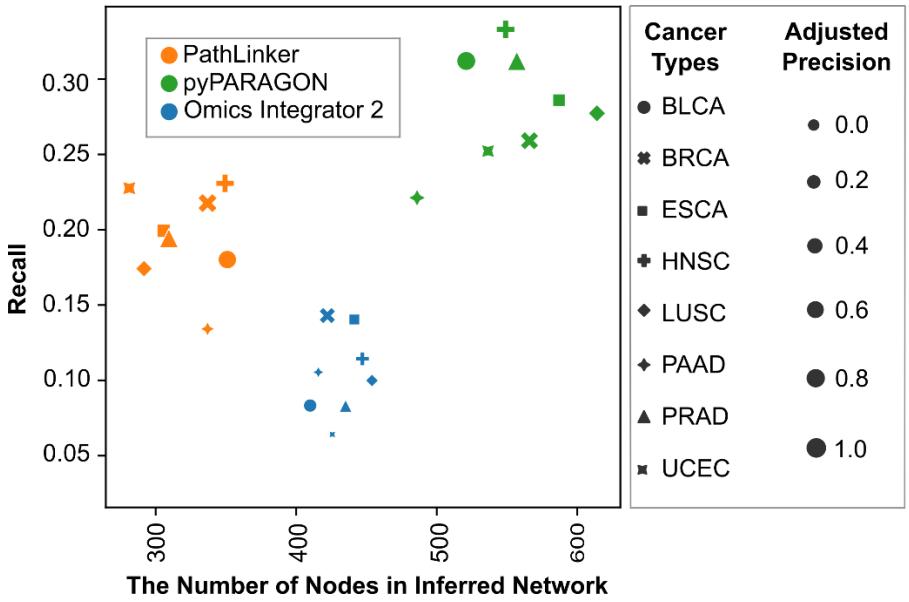

Figure 16: Performance of contextualized cancer-specific networks for eight distinct cancer types. Marker size represents precision, while recall and network sizes are shown on the x-axis and y-axis. The recall score represents the ratio of correctly predicted cancer driver genes in cancer-specific networks to total number of drivers. pyPARAGON achieved better recall scores for each cancer type without having a decrease in precision scores.

4.2.4. Running time analysis of pyPARAGON

During running-time analysis, we infer context-specific networks by randomly selecting initial networks and creating reference networks. We figured out context-specific networks by using random starting node sets with 25 to 750 nodes spread out over a network with 15,000 nodes and 544,249 interactions. Figure X demonstrates the duration of the inference series. We used 100 random initial nodes to infer context-specific networks over random reference networks, which are composed of 1,000 to 20,000 nodes, in the second running time measurement. Their duration is drawn in Figure x. We have nested functions to determine graphlets. Initially, pyPARAGON identifies directly interacting nodes among node pairs. Then, 3-node graphlets are specified based on the interaction knowledge of 2 nodes. Similarly, looking at neighbors of known 3-node graphlets, pyPARAGON determined 4-node graphlets. Therefore, due to these two nesting processes in graphlet determination, quadric running time was observed in Figures X and Y with $O(n^2)$.

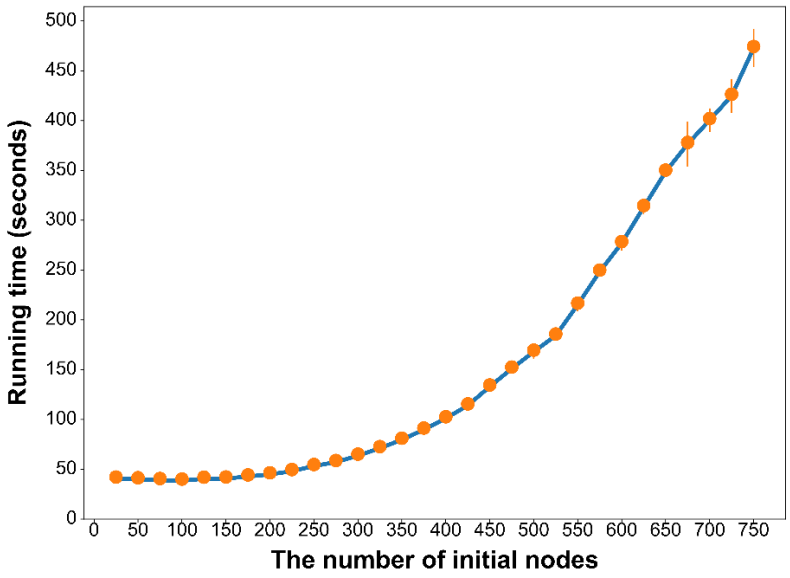

Figure 17: Running time graph based on the initial node size. The randomly selected node sets composed of between 25 and 750 nodes are recruited for the construction of context-specific networks over the random network with 15 000 nodes. Running time exponentially increases depending on the number of initial nodes.

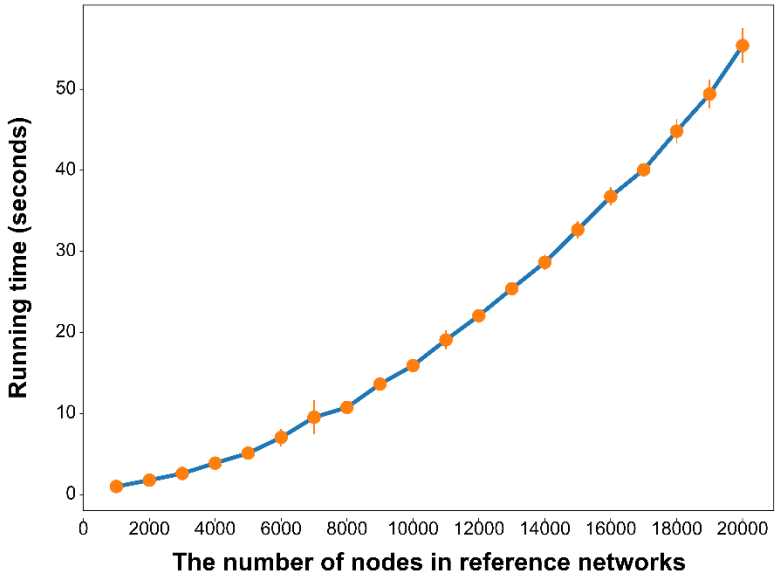

Figure 18: Running time graph based on the network size. Using 100 randomly selected initial nodes, we constructed context-specific networks on different size of reference networks between 1000 and 20000. Running time exponentially increases depending on the number of nodes in the reference networks.

CHAPTER 5

IMPLEMENTATION OF pyPARAGON

The manifestation of a disease phenotype often represents the complex interactions of many pathobiological processes within a complex network[240]. The high number of interactions within the human protein interaction network (interactome) suggests that complex diseases cannot be regarded as independent of each other at the molecular level due to the presence of many genetic and environmental factors[143]. Recently, advancements in high-throughput methods, data mining, and bioinformatics methodologies have enabled the exploration of human disorders at different molecular levels. Tools integrating multi-omics datasets may provide valuable insights into the molecular structure of diseases, leading to better knowledge of disease correlation. However, these tools have inherent challenges in identifying their significant genetic factors[95]. Network-based models provide contextualized solutions covering different data types for various tasks, such as the identification of novel disease proteins, drug targets, patient stratification, and functional modules in diseases[132,241]. Small communities in networks can be more robust and reliable compared to individual biomarker genes based on patterns in omics and may attain superior precision in categorizing diseases[242]. Thus, we contextualized cancer and neurodevelopmental disorders using various omics datasets. In the first case study, we implemented pyPARAGON to construct tumor-specific networks. In the downstream analysis, we utilized these contextualized networks to stratify patients and identify their functional modules and specific drugs. In the second case study, we compared cancer and neurodevelopmental disorders by modeling breast cancer and autism spectrum disorder (ASD) to reveal their common pathways and differences in signal strength.

5.1. Methods:

5.1.1. Case Study 1: Contextualization of breast cancer samples

5.1.1.1. Data preparation

We obtained phosphoproteomics datasets for 105 breast cancer patients and three healthy samples[243]. We considered only phosphosites observed in at least half of the samples to eliminate noisy signals and had a standard deviation larger than 0.5 in normalized data. Followingly, the remaining phosphosites were classified into two criteria: i) A log-2-fold-change (LFC) greater than 2; ii) highly or less phosphorylated in the Gaussian Mixture Model (GMM)[244]. Within the framework of GMM, we divided phosphoproteomics into three divisions: highly, less, or normally phosphorylated. We randomly conducted 100 iterations of GMM for each patient. Subsequently, we selected proteins with high or low phosphorylation levels in 95% of models. We weighed differential phosphoproteins ranging from 0.5 to 1.

5.1.1.2. Construction of tumor-specific networks

During the contextualization with the tumor-specific network, we implemented pyPARAGON, detailed in Chapter 4, by assigning the selected, significant phosphoproteins as a scored seed node set for each sample (**Figure 17**). As a reference interactome, we recruited HIPPIE v2.3[8] after removing self-interactions. We kept confidence scores of interactions without filtration. Additionally, we selected the highest scores for repeated interactions. As an example of supervised usage of pyPARAGON, we assigned significantly frequent graphlets G_2, G_5, G_6, G_7, and G_8, which had been identified for pathway reconstruction in Chapter 4. We used the following parameter sets: damping factor (λ) and flux threshold (τ) are set to 0.5, and the maximum number of interactions is 2000.

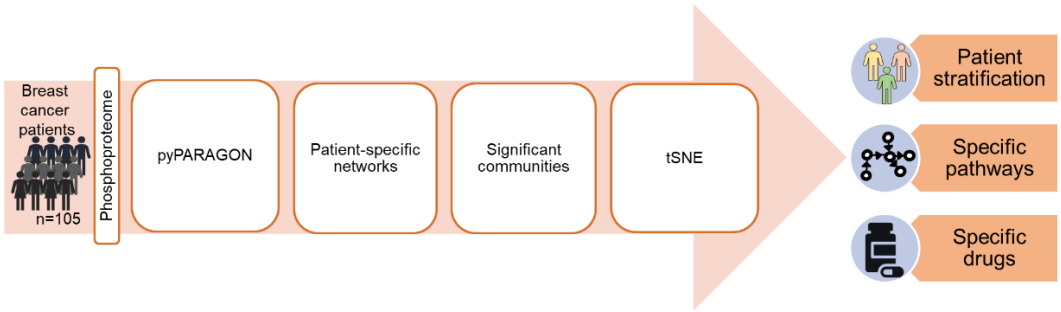

Figure 19: Contextualization and downstream analysis of tumor-specific networks. pyPARAGON used seed node sets from phosphoproteomics datasets of 105 breast cancer patients to construct tumor-specific networks. We extracted significant communities associated with gene ontology annotations. Then, we transformed network knowledge into the similarity matrix via significant communities to stratify patients. Also, specific pathways and drugs are determined using significant communities.

5.1.1.3. Identification of functional communities and downstream analysis

To identify functional communities, we divided tumor-specific networks into communities through the Louvain community detection methods[151]. We tested communities with the hypergeometric text[155] described in Chapter 4. As prior knowledge, we obtained biological processes from Gene Ontology (GO) annotations[81].

To transfer information within the contextualized network to the vector space, we scored "1" for annotations represented with functional communities and "0" for unrepresented annotations. Then, t-distributed stochastic neighbor embedding (t-SNE) algorithm was implemented to reduce the vector space matrix to two components[245]. Then, we clustered the patients with agglomerative clustering through the Euclidean distance. Also, we calculated the similarity matrix through the pairwise cosine similarities between the enriched biological processes of all paired patients by applying **Formula 5.1**. Followingly, we construct the patient-patient similarity network by adding an edge between patients with similarity scores greater than 0.5.

$$Sim_{Cos} = \frac{A.B}{||A||\,|||B||} \quad (5.1)$$

For biological interpretation, we analyzed the survival probabilities of patients after identifying functional communities associated with the specific biological processes from GO annotations[81]. Then, we detected specific pathways in cellular processing and

signal transduction from KEGG[70], and therapeutic drugs from the therapeutic target database[246]

5.1.2. Case study 2: Contextualization of neurodevelopmental disorders and cancers

5.1.2.1. Data preparation

We constructed a network model of ASD as an example of neurodevelopmental disorders (NDD), while breast cancer is an example of cancer (**Figure 18**). mutations were retrieved from denovo-db[28], NDD composed of human germlines de novo variants of 20 phenotypes. ASD dataset in denovo-db has been composed of targeted sequencing of 3203 patients, coming from either whole exome or whole genome studies. We only considered the point mutations affecting the canonical protein structures. Therefore, we initially mapped the genomic coordinates of mutations into protein structures using VarMap[247]. In this way, we identified 4881 unique mutations on 3839 genes. We assigned 190 genes to the seed node set by selecting genes seen in at least 3 patients. We retrieved cancer driver mutation, tabulated in the Catalog of Validated Oncogenic Mutations on the Cancer Genome Interpreter (CGI)[248]. We only considered 3688 driver mutations on 237 including missense or nonsense mutations. 168 genes associated with breast cancer were recruited to the seed nodes.

5.1.1.1. Construction of disease-specific networks

We implemented pyPARAGON by using HIPPIE v2.3 as a reference interactome[8]. To use our novel tool in an unsupervised manner, the union of all graphlets constructs GGN without texting the significance of any graphlets. In the following steps, we set the damping factor ($\boldsymbol{\lambda}$) as 0.5 for propagation and the flux threshold ($\boldsymbol{\tau}$) as 0.8 to select the interaction with the highest scores.

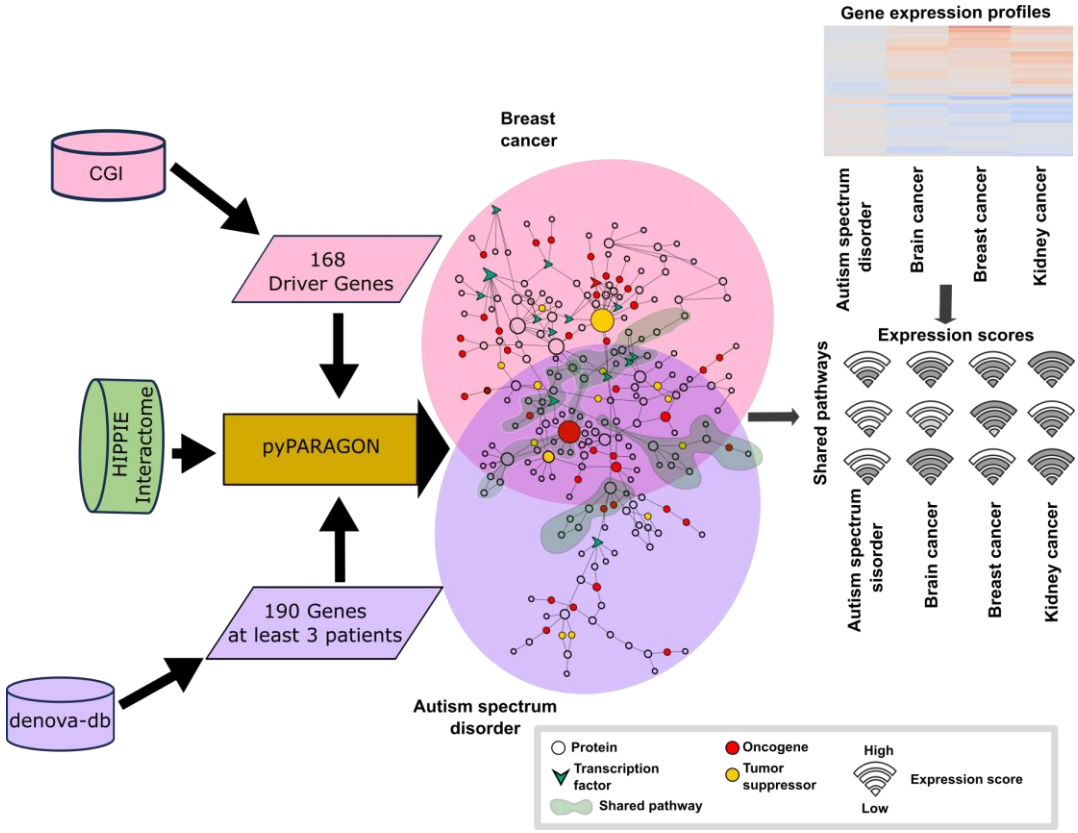

Figure 20: A conceptual representation of network comparison analysis between NDDs and cancer. Two distinct networks (left panel) were reconstructed for breast cancer (large pink circle) and ASD (large purple circle). These two networks have both shared (shaded green) and separated regions. These networks contain oncogenes (red circle), tumor suppressors (yellow circle), and TFs (green V-shapes). The transcriptome analysis (upper-right panel) associates the expression levels of the nodes with the pathway activity. Each enriched pathway in the network can be quantified with the average expression level of its nodes, which is called "pathway scoring." The score of each shared pathway (1, 2, ..., n) for each disease (ASD, purple; cancer, red) is calculated (shown as a Wi-fi icon where the higher score is the stronger signal).

5.1.1.2. Identification of the common pathways

By examining transcription factors (TFs), target genes, and the pathways connecting both diseases, we were able to deduce the shared functions of disease networks in overlapping network areas. TFs and their targets, obtained from TRRUST v2[64], were mapped into disease-specific networks. TFs in these networks were referred to specific transcription factors (STFs). Genes targeted by STFs were determined as regulated genes by disease-specific networks. We conducted the overrepresentation analysis for the commonly regulated genes among ASD and breast cancer to identify their shared pathways. The overrepresentation tool, WebGestalt[249] was recruited with KEGG[70], and Reactome[250] databases ($p < 0.05$ and FDR < 0.05)

5.1.1.3. Assessments of pathways

We used processed RNA expression data from samples of ASD, breast, kidney, and brain cancer, which are reported in **Table 4**[251]. The ASD dataset included combined data from three investigations, namely from frontal brain samples. It encompassed a total of 34 samples from individuals with ASD and 130 samples from control subjects. We used comprehensive datasets for breast, kidney, and brain tumors, consisting of 7, 10, and 8 trials, respectively. 3579 genes were identified as differentially expressed in ASD populations, whereas 11,629 genes were identified as differentially expressed in cancer cohorts, using z-scores.

Table 4: Expression profiles of diseases

Phenotype	Cases	Control	Datasets
ASD	34	130	GSE28475, GSE28521, (Gupta et al. 2014).
Brain cancer	942	104	GSE4290, GSE9385, GSE74195, GSE68848, GSE15824, GSE42656, GSE44971, GSE50161
Breast cancer	1494	249	GSE10810, GSE31448, GSE42568, GSE54002, GSE65216, GSE45827, GSE29431
Kidney cancer	400	266	GSE11151, GSE77199, GSE47032, GSE53757, GSE53000, GSE66272, GSE68417, GSE71963, GSE40435, GSE7635

The signal strength and mutation vulnerability of the common pathways were used as pathway assessment metrics. The signal deviation is measured considering the expression level of each gene in the given pathway. To determine the expression score (ES) of a particular pathway (P), we computed the mean absolute signal differences of pathway[252–254] using **Formula 5.2**. The given pathway, P=(G, E, U), consisting of genes/proteins (g_1, g_2, …, g_n, ∋ G), expression of genes ($|e_1|$, $|e_2|$, …, $|e_n|$ ∋ E), and the number of unique mutations belonging to genes (u_1, u_2, …, u_n ∋ U). We assessed the mutation vulnerability of a pathway by determining the propensity score (PS) of the pathway based on the number of unique mutations by using **Formula 5.3**.

$$ES_p = \frac{\sum_{k=1}^{n} |e_k|}{n} \quad (5.2)$$

$$PS_p = \frac{\sum_{k=1}^{n} u_k}{n} \quad (5.3)$$

5.2. Results:

5.2.1. Case Study 1: Tumor-specific network inference reveals hidden commonalities across tumors

We constructed tumor-specific networks of 105 breast cancer patients[243] by assigning significant phosphoproteins as seed nodes to pyPARAGON. Constructed networks are divided into functional subunits of networks regarded as modules or communities. These communities were annotated with associated biological processes and KEGG pathways. pyPARAGON applies hypergeometric tests to detect active modules with a statistically significant overrepresentation in various biological processes. The tumor-specific network shown in **Figure 19** is an example of an active module network with strong KEGG pathway associations.

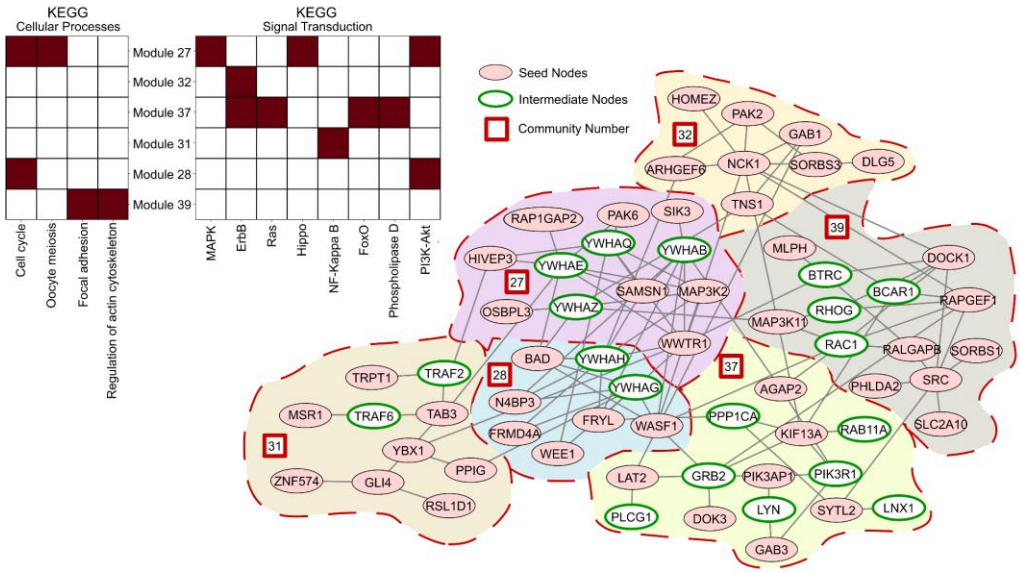

Figure 21: Example of active modules in a tumor-specific network constructed by pyPARAGON (TCGA-A8-A079). Significantly phosphorylated proteins were used as the initial (seed) node-set (colored pink), and intermediate nodes predicted by pyPARAGON are in green circles. Active modules are associated with at least one significantly overrepresented KEGG pathway bordered with dashed red lines and numbered within red boxes. The pathways belonging to cellular processes and signal transduction are listed in the top left chart.

We converted tumor-specific networks into a vector space by tagging community knowledge with biological processes in GOA. The patients were eventually categorized into four categories by using t-distributed stochastic neighbor embedding (t-SNE) reducing the matrix of biological processes. (**Figure 20A**). The 20 most

frequently identified biological functions for each cluster were listed in **Appendix B**. Critical biological processes observed in at least two clusters have been used for biological interpretations (**Figure 20B**). The ubiquitin-dependent protein catabolic process is the most often linked biological process in patient Cluster-1, characterized by the presence of many transcription factors and enzymes. Ubiquitination, a post-translational modification, is a complex enzymatic process that plays a role in regulating cancer metabolism[255]. Cluster-2 patients often have a shared occurrence of the mitotic cytokinesis process. Cytokinesis abnormalities lead to a rise in chromosomal instability, vast genomic alterations, and point mutations, promoting intra-tumoral heterogeneity[256,257]. The patient similarity network (**Figure 20C**) reveals that, as a result of heterogeneity, only five patients exhibit similarity scores over 0.5. Remarkably, we discovered that the process of nervous system development (NSD) was the most prevalent biological process in Cluster-3. Lung cancer is the primary cause of central nervous system metastases, followed by breast cancer[258]. Within our datasets, only two patients exhibited metastases. We could identify two cases exhibiting the NSD process inside Cluster-3. Cluster-4 had regulatory processes involving the structure of the actin cytoskeleton that were relevant to the onset, progression, and treatment of cancer. Rho GTPases, which belong to the Ras GTPase superfamily, have a crucial function in this regulation[259]. Our analysis revealed that patients belonging to Cluster-4 have a significantly lower likelihood of survival compared to those in Cluster-1 (**Figure 21A**).

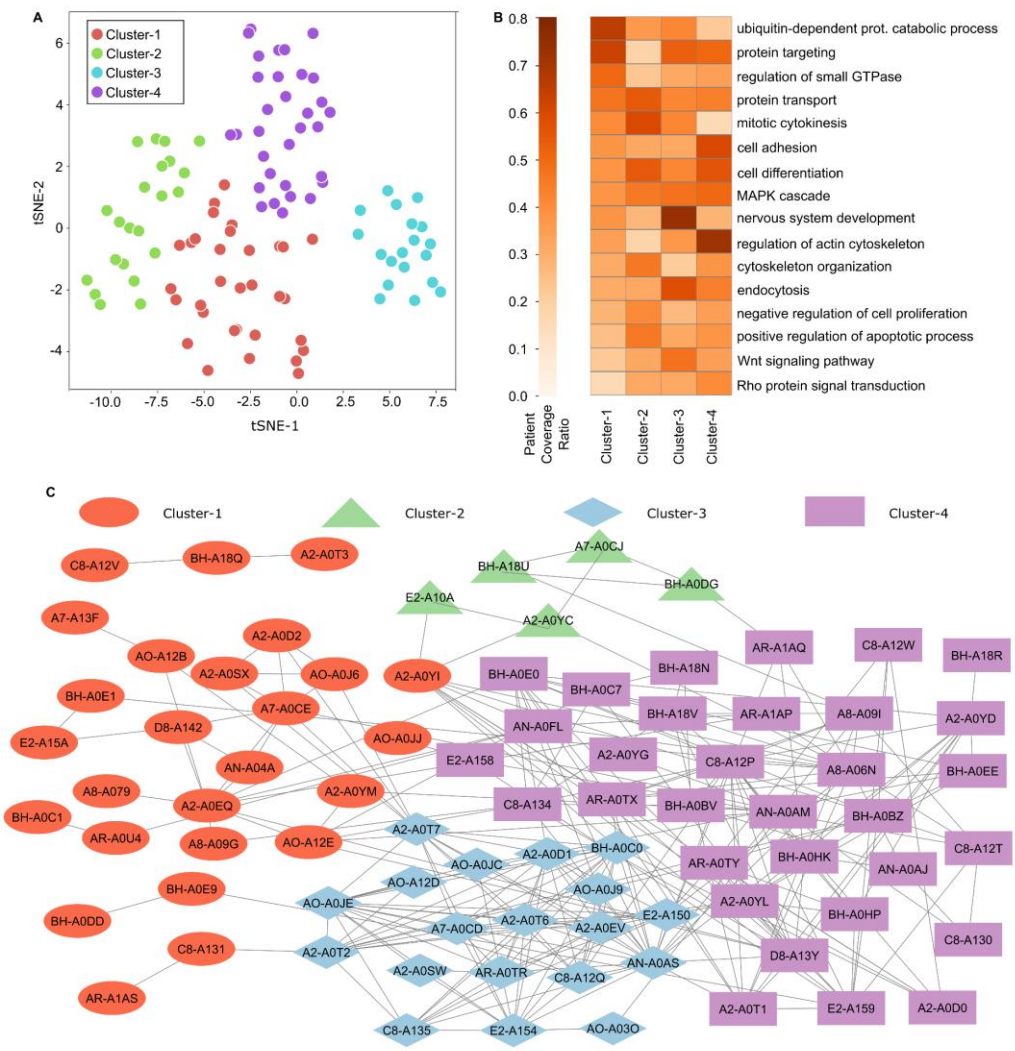

Figure 22: Stratification of tumors and associated biological processes with patient clusters. **A)** 105 breast cancer tumors are stratified into four clusters based on significant biological processes in their network modules: Cluster-1 (32 patients), Cluster-2 (22 patients), Cluster-3 (19 patients), and Cluster-4 (32 patients). **B)** Heatmap of patient coverage ratio for each cluster and significant process pairs. A biological process is included in the heatmap if it is enriched in at least two clusters. The patient coverage ratio represents patients with the enriched biological process in the corresponding clusters. **C)** The similarities of 105 patients were calculated through a cosine similarity score of meaningful biological processes between patient pairs. In the similarity network, we illustrated interactions between patients with similarity scores greater than 0.5 (82 patients, 262 interactions). In the similarity network, we displayed interactions between patients with similarity scores greater than 0.5. (82 patients, 262 interactions). Using the t-SNE algorithm and agglomerative clustering, we divided the patients into four groups based on biological processes. Cluster-1 was represented by red ellipses, Cluster-2 by green triangles, Cluster-3 by blue diamonds, and Cluster-4 by purple rectangles. Most patients in Cluster-2 do not have obvious similarities in patient pairs, while most in Cluster-3 and Cluster-4 do have higher similarities and more interactions in the patient similarity network.

We also utilized KEGG pathway information to identify their associated modules and overrepresented pathways in these clusters (**Figure 21B**). The cell cycle and PI3K/Akt signaling pathways are prevalent and often seen in clusters, except for Cluster-2. These pathways are more frequently observed in Cluster-1 compared to Cluster-4. The Ras signaling pathway plays a crucial role in drug resistance owing to the bypassing of drug action mechanisms in the signaling network[260,261]. The module related to the Ras-signaling pathway is illustrated in **Figure 21C**. In this module, pyPARAGON connected phosphoproteins with intermediate nodes such as KRAS, NRAS, HRAS, RHOA, and RHOD.

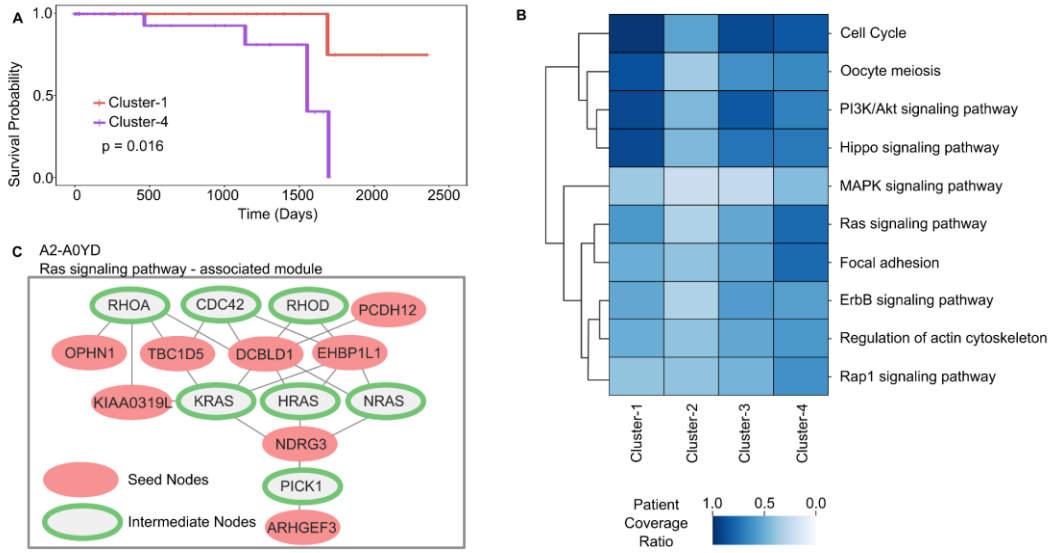

Figure 23: Survival analysis and cluster-specific KEGG pathways **A)** Kaplan-Meier analysis shows the survival probabilities of Cluster-1 (red) and Cluster-4 (purple). **B)** Heatmap shows significantly enriched KEGG pathways in active modules. **C.** The example module of A2-A0YD network corresponding to the Ras signaling pathway is shown where seed nodes are red and intermediate nodes are green.

For 105 breast cancer patients, we regained 8297 drugs and 330 therapeutic targets from the Therapeutic Target Database[246]. Additionally, we identified active modules related to 161 pathways. **Figure 22** illustrates an example of context-specific drugs for the active modules of patient A2-A0YD. Adagrasib (MRTX849) and Sotorasib particularly inhibit the Ras signaling-associated module. Both drugs are newly developed inhibitors of the KRASG12C protein the FDA autherized[260,262].

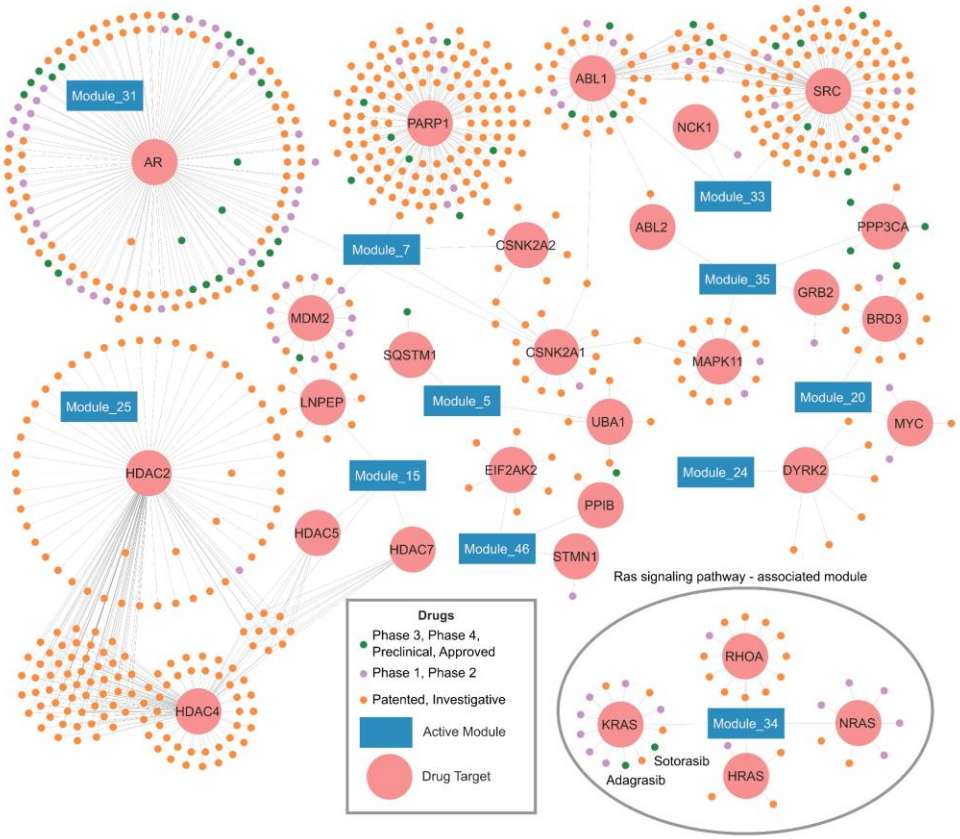

Figure 24: Drug-module interaction network of a patient (TCGA-A2-A9YD). Drugs are shown in three colors corresponding to three categories: drugs in phase 3, 4, or preclinical stage, and authorized drugs in green; drugs in phase 2 or 3 in purple and patented and investigational drugs in pink.

5.2.2. Case Study 2: Disease-specific networks identifies shared pathways

5.2.2.1. Different TFs regulate the shared pathways in ASD and breast cancer

In order to unravel the genetic associations and distinctions between neurodevelopmental disorders (NDDs) and cancer; we first used accessible mutation databases such as denovo-db[28] and CGI[248]. Denovo-db contains de novo mutation profiles, including neurodevelopmental disorders (NDDs) and other illnesses, for 9,736 samples. On the other hand, CGI covers the catalog of validated oncogenic mutations, including oncogenes and tumor suppressors. We constructed ASD- and breast cancer-specific networks using frequently mutated genes, driver genes, and HIPPIE interactome. The ASD-specific network contains 350 proteins and 1291 interactions, while the breast cancer-specific network has 284 proteins and 1878 interactions. Some crucial transcription factors (TFs) harboring cancer driver mutations, such as Myc, p53, and Jun, are not frequently mutated in ASD. On the other

hand, rewiring the signaling network allows mutated genes to control these TFs in the ASD-specific network indirectly. 23 transcription factors were common in both the ASD-specific and breast cancer-specific networks (**Figure 23A**). TF complexes such as Myc/Max or Jun/Fos (also known as AP-1, activator protein 1) regulate the expression of multiple genes that are part of the MAPK phosphorylation cascade in signal transduction[263,264]. Complexes composed of common TFs primarily play key roles in cell cycle regulation through their targets, such as E2F mediating cyclin-dependent kinases (CDKs) in cell proliferation[265,266]. The transcription factors (TFs) present in both ASD- and breast cancer-specific networks together control the expression of 752 genes that these TFs target. The disease models in both networks can follow distinct wiring strategies to regulate common pathways since various transcription factors govern the transcription of the same genes. Overrepresentation analysis revealed that many transcription factors control similar pathways, such as p53, FOXO, PI3K/AKT, MAPK, and JAK/STAT signaling pathways (**Figure 23B**).

5.2.2.2. Gene expression and signaling strength of the shared pathways

After constructing the networks and identifying the TFs and their targets, our analysis focused on the signal levels in these networks by comparing the differential gene expressions between healthy and disease samples. Due to the presence of multiple molecular functionalities, it is difficult to ascertain the impact of this signal modification on these shared pathways. Therefore, we calculated the mean absolute values of the differential expression of the participants of the given pathway, which is the expression score for the pathway. This score serves as a measure of the signal strength within these pathways. The expression scores of the overrepresented pathways revealed that ASD exhibited considerably reduced signal intensity compared to breast, brain, and kidney malignancies (**Figure 23C**), affecting the cell cycle during the G1 phase. Stimulus and feedback loops are responsible for regulating the strength and duration of signaling[267]. Excessive expression and various combinations of mutations in these pathways impair cellular functions and may control the development and onset of diseases.

The expression pattern of ASD in common biological pathways highlights differentiation. Cell differentiation rapidly decreases the multiplication ability of cells

and enhances their ability to withstand cancer-causing mutations[268]. ASD mutations mostly occur during embryonic development and do not progressively accumulate like cancer mutations. The propensity score of pathways, which represents the likelihood of mutations occurring on a gene in a specific pathway, indicates that cancer mutations tend to accumulate within specific pathways. Shared pathways in ASD have low propensity scores (**Figure 23C**). ASD individuals, due to their preexisting mutational load, are more prone to developing multifactorial and/or polygenic disorders such as cancer[269,270]. Simultaneously, their weak/moderate effect might induce cell cycle arrest and affect cellular differentiation capacities.

5.2.2.3. TFs highlight patterns of differentiation in NDDs and proliferation in cancer.

We compared the expression patterns of ASD and breast cancer patients using 71 transcription factors that regulate similar pathways. Our observation reveals that 57 individuals exhibit the expression score in ASD, whereas 21 transcription factors (TFs) have unique expression patterns in both ASD and breast cancer. These TFs are categorized into three separate categories. Cluster-1 and Cluster-2 exhibited a significant distinction, but Cluster-3 included genes that did not exhibit a discernible variation in the heatmap of gene expressions (**Figure 24A**). MCM2, STAT1, BRCA1, and MCM5 in Cluster-1 exhibit overexpression in the cancer samples. These genes primarily contribute to cell proliferation, and their overexpression profiles in cancer stimulate cell division and growth[271,272]. In contrast, ASD samples have comparatively lower expression levels of transcription factors regulating cell growth. STAT1, serving as a tumor suppressor and an oncogene in cancer, has a dual function in both cellular differentiation and proliferation. JUN, SMAD3, SMAD4, and KLF2, which are part of Cluster-2, are involved in the process of cell differentiation[273–276]. Their moderate levels of expression in ASD indicate that they are able to sustain the cellular differentiation stage. In order to elucidate the signal flow originating from these TFs, we established regulatory interaction within shared pathways by identifying the target genes controlled by these TFs. We expanded the analysis to include the regulatory interactions between targeted TFs and their corresponding target genes since TF may also regulate other TFs within the same pathway (**Figure 24B**). The expression patterns of differentiation and proliferation in individuals with ASD show a modest

level, indicating a mild activation of cell proliferation signals[277]. Nevertheless, the inhibition of cell differentiation and the overexpression of proliferation suggest a robust stimulation of the cell division process in cancer.

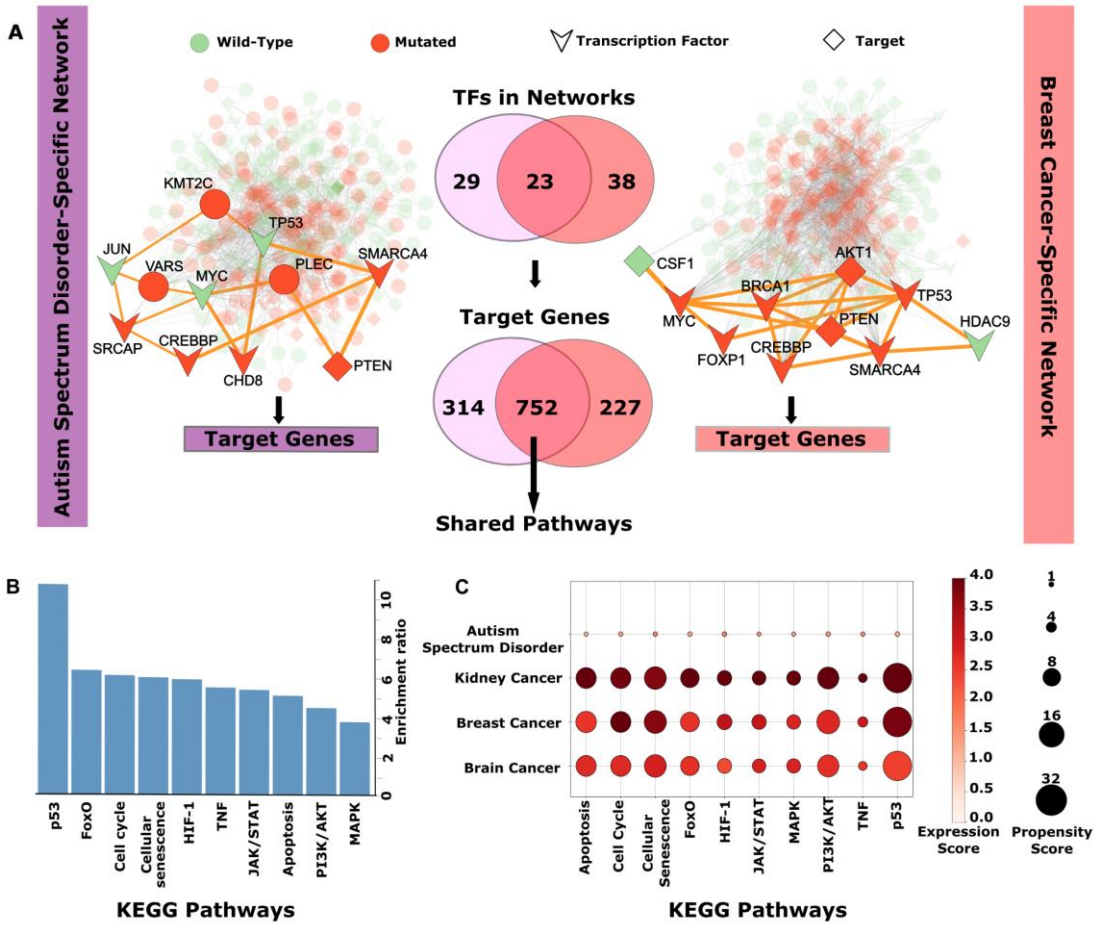

Figure 25: ASD- and breast cancer-specific networks regulating common pathways. **A)** Disease-specific network reconstruction for ASD and breast cancer is performed by using pyPARAGON tool, where the frequently mutated genes are used as seeds. The nodes in reconstructed networks involve wild type (green circle), mutated genes (red circle), TFs (chevron), and TF-targets (diamond). The complete ASD-specific network (left side) features the mutated proteins (SRCAP, BRG1, PTEN, etc.) in ASD cases and reveals disease-associated proteins (Jun, p53, and Myc). The breast cancer-specific network (right side) illustrates driver genes, although some driver genes, such as TP53 and MYC, are not frequently mutated in ASD. Both ASD- and breast cancer-specific networks involve 23 common TFs targeting 752 common genes. These common targets are employed to identify shared pathways. **B)** Overrepresentation analysis determines significant shared pathways (FDR $\leq$ 0.05) related to cell differentiation and proliferation among KEGG pathways. The pathways include MAPK, PI3K/AKT, and JAK/STAT. These shared TF-target genes play a significant role in cell fate by altering the signal strength and flow, as well as cell cycle and cellular senescence. HIF-1 hypoxia-inducible factor 1, TNF tumor necrosis factor. **C)** Signal changes in shared pathways are illustrated with the expression scores of pathways, the mean of the absolute z-scores of proteins in a given pathway. We define expression

scores as a mean of the absolute z-scores of proteins in a given pathway to indicate the magnitude of the deviation from the average expression values of the normal samples, regardless of the direction of the change. The vulnerability of common pathways to mutation is measured with a propensity score, the average unique mutation in the pathway. The darker red represents a higher change in expression scores of genes in the pathway, and the larger circle shows a higher mutation propensity for the corresponding pathway. ASD has the most minor signal differences and mutation propensities compared to all cancer types in shared pathways, where kidney cancer has the highest signal difference. However, there is an insignificant difference in mutation propensities among cancer types. The higher expression scores in cancer types point to stronger signal changes in pathways critical for cell fate, such as proliferation and differentiation. The higher propensity scores in cancer reveal that cancer mutations tend to group in shared pathways. Thus, shared pathways are more vulnerable to cancer than ones in ASD. However, mutation loads and signal deviations on the shared pathways might make ASD patients more fragile to cancer onset.

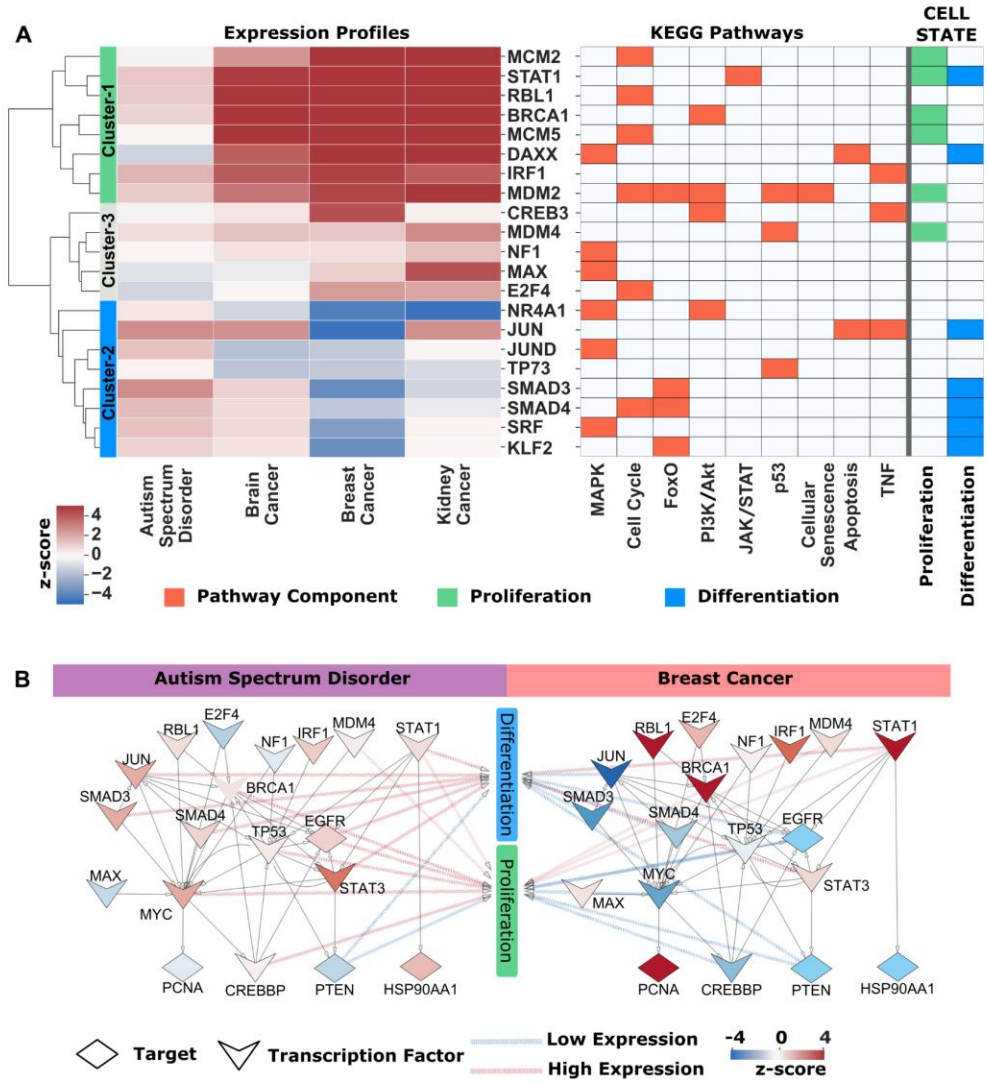

Figure 26: Differential TFs drive to proliferation in cancer and differentiation in ASD **A)** 21 TFs were identified to be at least one time differentially expressed more (less) in ASD than in other cancer types. On the left hand, the heatmap of these differentially expressed genes (high in red, low in blue) clustered expression z-scores into three groups. On the right hand, the pathways TFs belong to, and related cell states (proliferation, green; differentiation, blue) are demonstrated. Genes more expressed in cancer types than in ASD mainly belong to the proliferation state, while genes related to differentiation are predominantly more expressed in ASD than in cancer types. **B)** Differences between proliferation and differentiation on shared pathways. The signal flows from TFs (chevron) to targets (diamond) in common parts of ASD- and breast cancer-specific networks and in shared pathways were demonstrated with z-scores. The low and high expression levels were illustrated with blue to red, respectively. The relationship between cell state and proteins is represented with arrows whose color also demonstrates the level of expressions, low or high. Differentiation-related proteins, such as Jun, SMAD3, and SMAD4, mainly have low expression profiles in breast cancer, while most are highly expressed in ASD. PTEN, EGFR, and STAT1, related to proliferation and differentiation, have similar expression profiles.

CHAPTER 6

DISCUSSION

Advancements in high-throughput omics technologies have driven complex data issues and integration challenges. A wide range of interactions has been systematically characterized in contextualized networks, including protein-protein interactions (PPIs), interactions between transcription factors and genes, and the effects of medicines and small molecules on gene expression. However, during the contextualization of omics datasets, network-based tools encounter several challenging issues: i) Sparse outputs of omics datasets in reference networks or prior knowledge are a source of missing essential points in networks. ii) Interpretation methods can miss hidden knowledge that connects significant hits in omics datasets while evaluating multi-omics datasets. iii) Well-studied proteins in reference networks come along with bias in contextualization. iv) Highly connected nodes, or hubs, bring about unspecific and noisy interactions in inferred networks. Firstly, we clarified these challenging issues by assessing reference networks and network reconstruction algorithms. Then, we developed and launched pyPARAGON (PAgeRAnk-flux on Graphlet-guided network for multi-Omics data integratioN), which combines network propagation with graphlets to integrate multi-omics data. pyPARAGON improves precision and reduces the presence of non-specific interactions in signaling networks by using network motifs[278].

6.1. Evaluation of the network reconstruction approaches on various interactomes

We comprehensively analyzed the characteristics of interactomes from various sources and evaluated the effectiveness of four network reconstruction strategies on established routes. PathwayCommons has the most extensive coverage of nodes and edges across all interactomes, including cancer driver genes and known pathways, since it has the maximum number of nodes and edges. However, the reconstruction

approaches using PathwayCommons exhibit notably lower accuracy values than the others. PathwayCommons shows a bias toward well-studied proteins due to the significant correlation between the degree and the number of publications of the nodes. Notably, HIPPIE and ConsensusPathDB have a similar bias, although their algorithms provide higher accuracy on these interactomes than PathwayCommons. The findings suggest that HIPPIE and ConsensusPathDB effectively reduce the impact of false positives while maintaining high confidence in the identified connections.

The all-pairs shortest path (APSP) algorithm yields the highest recall scores when accompanied by the highest false positive rate (FPR) score due to the inclusion of several false-positive edges and the true positives. Certain studies, like PathLinker[137], employ a distance threshold in calculating the shortest path, a restricted number of shortest paths between the source and target, or supplementary data that includes the orientation of the signal from the receptors to the transcription factors to control the rate of false positives. It is crucial to mention that we did not include any distance-based cutoff, supplementary data, or optimization in the APSP method. Consequently, the F1 and precision scores have an extremely low value in APSP. However, personalized PageRank with flux (PRF), heat diffusion with flux (HDF), and the prize-collecting Steiner Forest (PCSF) have comparable performance in terms of false positive and true positive edges. However, PRF, HDF, and PCSF have similar performance in terms of false positive and true positive edges. PCSF has the highest F1 score in comparison to PRF and HDF. Interactomes are an unbalanced dataset in which true-negative edges are much more than true-positive edges considering contextualization.

The Notch pathway reconstruction demonstrates that PCSF performs superiorly in identifying nodes with weak connections. However, PCSF should have been more comprehensive in uncovering the hidden nodes and their connections that facilitate communication between the Notch pathway and the PI3K-AKT-mTOR and JAK-STAT signaling pathways. Furthermore, the hidden nodes that connect signaling pathways are unable to form entirely true connections. While our study focuses on proteins as nodes, it is worth noting that pathways may also include small molecules and non-peptide nodes. Thus, the reconstruction algorithms are likely to introduce

false edges to include proper terminals. The lack of some nodes in the reference interactomes might contribute to the poor precision ratings.

The topological features and edge weights of the reference interactomes significantly impact network reconstruction strategies[279,280]. The presence of sparse data, where the number of edges in the target pathway is much lower than in the rest of the interactome, leads to high recall and poor precision scores[281]. Reconstruction algorithms of human signaling networks often exhibit a combination of poor precision and intermediate recall scores[137,282,283]. A further issue is that the node-based performance of reconstruction algorithms is superior to their edge-based performance. We have also seen the same trend in the results of our assessment.

Various tools use the topological characteristics of reference interactomes to predict new connections and eliminate false positive interactions[284–287]. Moreover, missing protein associations were identified using functional annotations, protein structures, and domain-domain interactions[241,288–290]. It is crucial to mention that we did not apply any additional method to filter out or modify interactomes[291] and the techniques to predict regulatory networks[292,293], during our assessment. The performance of the APSP, HDF, PRF, and PCSF algorithms may be affected by any modification or improvement made to the reference interactomes. The reference interactomes are undirected graphs, whereas signaling pathways are inherently directed ones. The directionality of the edges can be integrated using either the known or expected ones. Moreover, using the previously guided network may be an effective approach to disease modeling, especially in dealing with the complex structure of reference networks[36]. Thus, in the following part of the thesis, we aimed to reduce reference interactomes by trimming unassociated nodes and edges to enhance the performance of algorithms. Ultimately, biomolecular interactions exhibit a wide range of variations in both time and space. Thus, network reconstruction algorithms may be improved to include biological annotations and temporal and spatial interactions of proteins.

6.2. pyPARAGON unveiled hidden knowledge by contextualizing networks

This thesis introduces pyPARAGON as a network-based multi-omics data integration tool that combines the most common graphlets covering omics hits and the

Personalized PageRank algorithm to construct context-specific networks. Problems with sparse data and the increasing complexity of reference network connections are a source of difficulty for network inference algorithms. pyPARAGON reduced the effect of noise caused by highly connected nodes in the reference networks. The construction of graphlet-guided networks (GGN) especially maintains the scale-free characteristics of biological networks while minimizing noise. Additionally, the PageRank flux computation prioritized edges and was effectively combined with GGNs to infer context-specific networks. By identifying driver genes, we have expanded the scope of the missing value issue in building cancer-specific networks. pyPARAGON constructed the network-based models of various cancer types that covered a more precise and higher number of cancer drivers. Moreover, pyPARAGON can include modules and different kinds of annotations, such as biological processes, pathways, and pharmacological information, by inferring context-specific networks using phosphoproteomics. The results suggest that pyPARAGON can predict cancer biomarkers, drivers, drugs, and therapeutic targets.

Recent network inference tools, such as belief propagation[294], random walks[35], the prize-collecting Steiner Forest[34], heat diffusion[43], and shortest path algorithms[137], are encountering a significant challenge due to the presence of missing interactions and highly connected nodes, or hubs, resulting from extended integrations in reference networks[95]. Here, graphlets were used in our methods for trimming networks in our approaches. During the performance assessment, we compared pyPARAGON with two widely used tools such as PathLinker[137] and Omics Integrator 2[34] by reconstructing pathways and contextualizing various cancer types. Hub proteins may cause noises in the inferred network with unrelated interactions[143]. The prize-collecting Steiner Forest algorithm penalizes nodes according to their interaction number. Similarly, the flux calculation in pyPARAGON serves as a countermeasure against the negative impact of hubs on scoring interactions. Thus, Omics Integrator 2 and pyPARAGON have superior performance in predicting interactions. On the other hand, highly connected nodes reduce the length of the shortest pathways between seed nodes so that PathLinker can prioritize hubs. Omics Integrator 2 instantly halts the propagation of the seed nodes in a large reference network. However, the pyPARAGON tool implements the PageRank algorithm propagating the seed nodes, regardless of the

GGN. Therefore, pyPARAGON enhances the inference of interactions and the propagation of seed nodes in the network.

Even though integrative methods like pyPARAGON have worked well, long-term problems with network-based omics data merging still need to be fixed. It is worth mentioning that the characteristics and coverage of reference networks[295] are crucial to network-based approaches, while reference interactomes still need to be completed[296]. Due to incomplete knowledge of large reference interactomes, protein complexes have a higher tendency to construct a more significant number of topological modules than metabolic pathways[297]. Generic biological processes, like transcription and replication, can be found more often in contextualized networks. Therefore, due to the causal relationships, modular structures, and biological processes that are part of networks, it can also be challenging to interpret their meanings in a biological context. In addition, network-based approaches fail to evaluate the alternative copies of individual hits, such as diverse protein isoforms and post-translational modifications existing within the proteome. Although it provides more specialized functionalities, this information might be generic and obscured inside the network.

pyPARAGON only used graphlets consisting of interactions between 3 and 4 nodes rather than direct interactions of 2 nodes. Graphlet information, including graphlet degree distribution, graphlet frequencies, and probabilistic graphlets, may be included in network inference algorithms or used for biological interpretations[231,298–300]. However, the use of graphlet characteristics will entail a high computational expense. Permutation-based approaches may be further used in further research, where only hypergeometric tests on communities enhance context specificity more precisely. These communities may be detailed with mechanistic and causal relations for downstream analysis.

Various molecular abnormalities, especially in complex diseases like cancer and neurodevelopmental disorders (NDDs), may lead to indistinguishable clinical symptoms[301,302]. We utilized omics data from breast cancer tumors in CPTAC[243] as a case study to infer tumor-specific networks in which interacting protein modules regulate different biological processes and pathways. Patients could be clustered

together based on the overrepresented biological processes that were shown to be more prevalent and functional communities. Our case studies show that functional communities that share common driver genes, which recruit specific proteins, facilitate interpretation and various biological processes. Thus, pyPARAGON is an influential tool for detecting disease-associated molecular alterations and driver networks.

For further analysis of the complex relationship between genotype and phenotype, we contextualized disease-specific networks for autism spectrum disorder (ASD) and breast cancer. We identified distinct protein-protein interactions (PPIs) inside common pathways that regulate the cell cycle. The rewired networks may account for the varying signal levels in common pathways between ASD and breast cancer. Under physiological conditions, the MAPK and PI3K/AKT/mTOR pathways have crosstalk to regulate the cell cycle through feedback loops to maintain several cell processes such as growth, division, differentiation, and apoptosis. In the cancer context, these pathways are frequently hyperactivated[303–305]. The PI3K/AKT pathway plays a crucial role in the first stages of embryonic development and in preserving the ability of stem cells to differentiate into various cell types by suppressing the MAPK proliferation pathway[306–309]. The mutations induce signaling perturbations that may be categorized as weak/moderate and significant signaling alterations, represented by ASD and breast cancer, respectively. Strong signals promote cell growth, whereas weak or moderate signals might cause cells to exit the cell cycle for differentiation[310].

We have developed a new tool called pyPARAGON using graphlets and network propagation to infer context-specific networks. It reduces the impact of noise caused by nodes with many connections, maintains the characteristics of a scale-free network, and incorporates network modules and biological annotations. We can use its contextualized networks to predict biomarkers, medications, and therapeutic targets particular to the given situation. The communities inside the network have the potential to be used in downstream analysis to find mechanistic molecular relationships in complicated and rare diseases. pyPARAGON can integrate large-scale omics data into static network models for patients or diseases. The next version of pyPARAGON will be an extension to integrate omics data at the single-cell level to elucidate cell-type specific interactions.

REFERENCES

1. Pérez-Ortín, J. E., Tordera, V. & Chávez, S. Homeostasis in the Central Dogma of molecular biology: the importance of mRNA instability. *RNA Biol.* **16**, 1659–1666 (2019).

2. Chang, H. Y. & Qi, L. S. Reversing the Central Dogma: RNA-guided control of DNA in epigenetics and genome editing. *Mol. Cell* **83**, 442–451 (2023).

3. Buescher, J. M. & Driggers, E. M. Integration of omics: more than the sum of its parts. *Cancer Metab. 2016 41* **4**, 1–8 (2016).

4. Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O. & Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* **19**, 3735–3746 (2021).

5. Noor, E., Cherkaoui, S. & Sauer, U. Biological insights through omics data integration. *Curr. Opin. Syst. Biol.* **15**, 39–47 (2019).

6. Agarwal, M., Adhil, M. & Talukder, A. K. Multi-omics multi-scale big data analytics for cancer genomics. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 9498 228–243 (Springer Verlag, 2015).

7. Dugourd, A. *et al.* Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol. Syst. Biol.* **17**, 1–17 (2021).

8. Alanis-Lobato, G., Andrade-Navarro, M. A. & Schaefer, M. H. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res.* **45**, (2017).

9. Kamburov, A. & Herwig, R. ConsensusPathDB 2022: molecular interactions update as a resource for network biology. *Nucleic Acids Res.* **50**, D587–D595 (2022).

10. Wiel, L. *et al.* MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. *Hum. Mutat.* **40**, 1030–1038 (2019).

11. Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* **139**, 1197–1207 (2020).

12. Karlsson, M. *et al.* A single–cell type transcriptomics map of human tissues.

Sci. Adv. **7**, (2021).

13. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347**, (2015).

14. Gao, Y. *et al.* TRmir: A Comprehensive Resource for Human Transcriptional Regulatory Information of MiRNAs. *Front. Genet.* **13**, 808950 (2022).

15. Ben Guebila, M. *et al.* GRAND: a database of gene regulatory network models across human conditions. *Nucleic Acids Res.* **50**, D610–D621 (2022).

16. Demir, E. *et al.* The BioPAX community standard for pathway data sharing. *Nature Biotechnology* vol. 28 935–942 (2010).

17. Paull, E. O. *et al.* Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* **29**, 2757–2764 (2013).

18. Badia-i-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).

19. Wang, Y., Yang, Y., Chen, S. & Wang, J. DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration. *Brief. Bioinform.* **00**, 1–10 (2021).

20. Nativio, R. *et al.* An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. *Nat. Genet. 2020 5210* **52**, 1024–1035 (2020).

21. Tomazou, M. *et al.* Multi-omics data integration and network-based analysis drives a multiplex drug repurposing approach to a shortlist of candidate drugs against COVID-19. *Brief. Bioinform.* (2021) doi:10.1093/bib/bbab114.

22. Kapoor, M. *et al.* Multi-omics integration analysis identifies novel genes for alcoholism with potential overlap with neurodegenerative diseases. *Nat. Commun. 2021 121* **12**, 1–12 (2021).

23. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet. 2013 4510* **45**, 1113–1120 (2013).

24. The International Cancer Genome Consortium. International network of cancer genome projects. *Nat. 2010 4647291* **464**, 993–998 (2010).

25. Rudnick, P. A. *et al.* A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline. *J. Proteome Res.* **15**, 1023–1032 (2016).

26. Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nat. 2018 5557696* **555**, 371–376 (2018).

27. Li, H. *et al.* The landscape of cancer cell line metabolism. *Nat. Med. 2019*

255 **25**, 850–860 (2019).

28. Turner, T. N. *et al.* denovo-db: a compendium of human de novo variants. *Nucleic Acids Res.* **45**, D804–D811 (2017).

29. Bersanelli, M. *et al.* Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinformatics* **17**, 167–177 (2016).

30. Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* **14**, 8124 (2018).

31. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet. 2015 162* **16**, 85–97 (2015).

32. Kim, W. & Haukap, L. NemoProfile as an efficient approach to network motif analysis with instance collection. *BMC Bioinformatics* **18**, 37–45 (2017).

33. Wu, C. *et al.* A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. *High-Throughput* **8**, (2019).

34. Tuncbag, N. *et al.* Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLoS Comput. Biol.* **12**, 1–18 (2016).

35. Ahmed, R., Baali, I., Erten, C., Hoxha, E. & Kazan, H. MEXCOwalk: Mutual exclusion and coverage based random walk to identify cancer modules. *Bioinformatics* **36**, 872–879 (2020).

36. Hristov, B. H., Chazelle, B. & Singh, M. uKIN Combines New and Prior Information with Guided Network Propagation to Accurately Identify Disease Genes. *Cell Syst.* **10**, 470-479.e3 (2020).

37. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).

38. Arici, M. K. & Tuncbag, N. Performance Assessment of the Network Reconstruction Approaches on Various Interactomes. *Front. Mol. Biosci.* **8**, 1–19 (2021).

39. Turner, B. *et al.* iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford).* **2010**, (2010).

40. Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M. & Wodak, S. J. Interaction databases on the same page. *Nat. Biotechnol. 2011 295* **29**, 391–393 (2011).

41. Ayar, E. S., Dadmand, S. & Tuncbag, N. Network Medicine : From Conceptual Frameworks to Applications and Future Trends. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* **9**, 374–381 (2023).

42. Silverbush, D. *et al.* Simultaneous Integration of Multi-omics Data Improves the Identification of Cancer Driver Modules. *Cell Syst.* **8**, 456-466.e5 (2019).

43. Reyna, M. A., Leiserson, M. D. M. & Raphael, B. J. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* **34**, i972–i980 (2018).

44. Ideker, T., Galitski, T. & Hood, L. A NEW APPROACH TO DECODING LIFE: Systems Biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2003).

45. Rigden, D. J., Xos´, X., Fernández, X. M. & Fernández, F. The 2022 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res.* **50**, D1–D10 (2022).

46. Silverman, E. K. *et al.* Molecular networks in Network Medicine: Development and applications. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* vol. 12 1489 (2020).

47. Koutrouli, M., Karatzas, E., Paez-Espino, D. & Pavlopoulos, G. A. A Guide to Conquer the Biological Network Era Using Graph Theory. *Front. Bioeng. Biotechnol.* **8**, 504360 (2020).

48. Jin, S., Zeng, X., Xia, F., Huang, W. & Liu, X. Application of deep learning methods in biological networks. *Brief. Bioinform.* **22**, 1902–1917 (2021).

49. Kamburov, A., Stelzl, U. & Herwig, R. IntScore: A web tool for confidence scoring of biological interactions. *Nucleic Acids Res.* **40**, (2012).

50. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).

51. Assenov, Y., Ramírez, F., Schelhorn, S. E. S. E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **24**, 282–284 (2008).

52. Elmezain, M., Othman, E. A. & Ibrahim, H. M. Temporal Degree-Degree and Closeness-Closeness: A New Centrality Metrics for Social Network Analysis. *Math. 2021, Vol. 9, Page 2850* **9**, 2850 (2021).

53. Fagny, M. *et al.* Identification of Key Tissue-Specific, Biological Processes by Integrating Enhancer Information in Maize Gene Regulatory Networks. *Front. Genet.* **11**, 606285 (2021).

54. Guan, Y. *et al.* Tissue-Specific Functional Networks for Prioritizing Phenotype and Disease Genes. *PLOS Comput. Biol.* **8**, e1002694 (2012).

55. Oh, M. *et al.* DRIM: A Web-Based System for Investigating Drug Response at the Molecular Level by Condition-Specific Multi-Omics Data Integration. *Front. Genet.* **0**, 1053 (2020).

56. Lee, J. H., Park, Y. R., Jung, M. & Lim, S. G. Gene regulatory network

analysis with drug sensitivity reveals synergistic effects of combinatory chemotherapy in gastric cancer. *Sci. Reports 2020 101* **10**, 1–10 (2020).

57. Hollander, M., Hamed, M., Helms, V. & Neininger, K. MutaNET: a tool for automated analysis of genomic mutations in gene regulatory networks. *Bioinformatics* **34**, 864–866 (2018).

58. Wang, L. *et al.* Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nat. Methods 2023 209* **20**, 1368–1378 (2023).

59. Chen, X. *et al.* A novel method of gene regulatory network structure inference from gene knock-out expression data. *Tsinghua Sci. Technol.* **24**, 446–455 (2019).

60. Xie, Y., Wang, R. & Zhu, J. Construction of breast cancer gene regulatory networks and drug target optimization. *Arch. Gynecol. Obstet.* **290**, 749–755 (2014).

61. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).

62. Janky, R. *et al.* iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLOS Comput. Biol.* **10**, e1003731 (2014).

63. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021).

64. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).

65. Santos-Zavaleta, A. *et al.* RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res.* **47**, D212–D220 (2019).

66. Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G. & Orengo, C. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Struct. Des.* **18**, 1233–1243 (2010).

67. Ni, D., Lu, S. & Zhang, J. Emerging roles of allosteric modulators in the regulation of protein-protein interactions (PPIs): A new paradigm for PPI drug discovery. *Med. Res. Rev.* **39**, 2314–2342 (2019).

68. Azeloglu, E. U. & Iyengar, R. Signaling Networks: Information Flow, Computation, and Decision Making. *Cold Spring Harb. Perspect. Biol.* **7**, a005934 (2015).

69. Young, H. L. *et al.* An adaptive signaling network in melanoma inflammatory niches confers tolerance to MAPK signaling inhibition. *J. Exp. Med.* **214**, 1691–1710

(2017).

70. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).

71. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).

72. Catozzi, S. *et al.* Reconstruction and analysis of a large - scale binary Ras - effector signaling network. *Cell Commun. Signal.* 1–19 (2022) doi:10.1186/s12964-022-00823-5.

73. Vavouraki, N. *et al.* Integrating protein networks and machine learning for disease stratification in the Hereditary Spastic Paraplegias. (2021) doi:10.1016/j.isci.2021.102484.

74. Wingo, T. S. *et al.* Shared mechanisms across the major psychiatric and neurodegenerative diseases. *Nat. Commun. 2022 131* **13**, 1–19 (2022).

75. Wang, J., Wen, N. F., Wang, C., Zhao, L. & Cheng, L. ELECTRA-DTA: a new compound-protein binding affinity prediction model based on the contextualized sequence encoding. *J. Cheminform.* **14**, 1–14 (2022).

76. Lin, Y., Yan, S., Chang, X., Qi, X. & Chi, X. The global integrative network: integration of signaling and metabolic pathways. *aBIOTECH* **3**, 281–291 (2022).

77. Buphamalai, P., Kokotovic, T., Nagy, V. & Menche, J. Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nat. Commun. 2021 121* **12**, 1–15 (2021).

78. Elmentaite, R., Domínguez Conde, C., Yang, L. & Teichmann, S. A. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nat. Rev. Genet. 2022 237* **23**, 395–410 (2022).

79. Quan, P. *et al.* Integrated network analysis identifying potential novel drug candidates and targets for Parkinson's disease. *Sci. Reports 2021 111* **11**, 1–9 (2021).

80. Treutlein, B. *et al.* Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nat. 2016 5347607* **534**, 391–395 (2016).

81. Carbon, S. *et al.* The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).

82. Abascal, F. *et al.* Perspectives on ENCODE. *Nat. 2020 5837818* **583**, 693–698 (2020).

83. Schriml, L. M. *et al.* The Human Disease Ontology 2022 update. *Nucleic Acids Res.* **50**, D1255–D1261 (2022).

84. Piñero, J., Saüch, J., Sanz, F. & Furlong, L. I. The DisGeNET cytoscape app:

Exploring and visualizing disease genomics data. *Comput. Struct. Biotechnol. J.* **19**, 2960–2967 (2021).

85. Klopfenstein, D. V *et al.* GOATOOLS: A Python library for Gene Ontology analyses OPEN. *Sci. Rep.* **8**, 1–17 (2018).

86. Sherman, B. T. *et al.* DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, W216–W221 (2022).

87. Mi, H. *et al.* Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat. Protoc. 2019 143* **14**, 703–721 (2019).

88. Bludau, I. & Aebersold, R. Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nat. Rev. Mol. Cell Biol. 2020 216* **21**, 327–340 (2020).

89. Velmeshev, D. *et al.* Single-cell genomics identifies cell type–specific molecular changes in autism. *Science (80-.).* **364**, 685–689 (2019).

90. Arfin, S. *et al.* Oxidative Stress in Cancer Cell Metabolism. *Antioxidants 2021, Vol. 10, Page 642* **10**, 642 (2021).

91. Jeffery, C. J. Protein moonlighting: What is it, and why is it important? *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 373 (2018).

92. Uribe, M. L., Marrocco, I. & Yarden, Y. EGFR in Cancer: Signaling Mechanisms, Drugs, and Acquired Resistance. *Cancers 2021, Vol. 13, Page 2748* **13**, 2748 (2021).

93. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 1–15 (2017).

94. Shilo, S., Rossman, H. & Segal, E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat. Med.* **26**, 29–38 (2020).

95. Demirel, H. C., Arici, M. K. & Tuncbag, N. Computational approaches leveraging integrated connections of multi-omic data toward clinical applications. *Mol. Omi.* (2021) doi:10.1039/d1mo00158b.

96. Chen, C. *et al.* Applications of multi-omics analysis in human diseases. *MedComm* **4**, e315 (2023).

97. Burgess, D. J. Reaching completion for GTEx. *Nat. Rev. Genet. 2020 2112* **21**, 717–717 (2020).

98. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).

99. Santiago-Rodriguez, T. M. & Hollister, E. B. Multi 'omic data integration: A review of concepts, considerations, and approaches. *Semin. Perinatol.* **45**, 151456 (2021).

100. Aydin, B., Caliskan, A. & Arga, K. Y. Overview of omics biomarkers in pituitary neuroendocrine tumors to design future diagnosis and treatment strategies. *EPMA J.* **12**, 383–401 (2021).

101. Huo, Z. *et al.* Two-Way Horizontal and Vertical Omics Integration for Disease Subtype Discovery. *Stat. Biosci. 2019 121* **12**, 1–22 (2019).

102. Mihaylov, I., Kańduła, M., Krachunov, M. & Vassilev, D. A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models. *Biol. Direct 2019 141* **14**, 1–17 (2019).

103. Malod-Dognin, N. *et al.* Towards a data-integrated cell. *Nat. Commun.* **10**, (2019).

104. Kim, S., Oesterreich, S., Kim, S., Park, Y. & Tseng, G. C. Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. *Biostatistics* **18**, 165–179 (2017).

105. Kim, M. & Tagkopoulos, I. Data integration and predictive modeling methods for multi-omics datasets. *Mol. Omi.* **14**, 8–25 (2018).

106. Mirza, B. *et al.* Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes 2019, Vol. 10, Page 87* **10**, 87 (2019).

107. Martini, P., Chiogna, M., Calura, E. & Romualdi, C. MOSClip: multi-omic and survival pathway analysis for the identification of survival associated gene and modules. *Nucleic Acids Res.* **47**, e80–e80 (2019).

108. Manna, S., Roy, I., Majumder, D., Banerjee, A. & Pati, S. K. Multiple Data Integration Using Joint Non-negative Matrix Factorization. 667–677 (2022) doi:10.1007/978-981-16-2543-5_57.

109. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.* **13**, e1005752 (2017).

110. Mo, Q. *et al.* A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* **19**, 71–86 (2018).

111. Yeger-Lotem, E. & Sharan, R. Human protein interaction networks across tissues and diseases. *Front. Genet.* **0**, 257 (2015).

112. Grigo, C. & Koutsourelakis, P.-S. Bayesian Model and Dimension Reduction for Uncertainty Propagation: Applications in Random Media. *https://doi.org/10.1137/17M1155867* **7**, 292–323 (2019).

113. Wang, T. *et al.* MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun. 2021 121* **12**, 1–13 (2021).

114. Rappoport, N. & Shamir, R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics* **35**, 3348–3356 (2019).

115. Veyel, D. *et al.* Biomarker discovery for chronic liver diseases by multi-omics – a preclinical case study. *Sci. Reports 2020 101* **10**, 1–14 (2020).

116. Yan, K. K., Zhao, H. & Pang, H. A comparison of graph- and kernel-based -omics data integration algorithms for classifying complex traits. *BMC Bioinformatics* **18**, 539 (2017).

117. Peng, C., Zheng, Y. & Huang, D. S. Capsule Network Based Modeling of Multi-omics Data for Discovery of Breast Cancer-Related Genes. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **17**, 1605–1612 (2020).

118. Song, Q. *et al.* SMGR: a joint statistical method for integrative analysis of single-cell multi-omics data. *NAR Genomics Bioinforma.* **4**, (2022).

119. Song, M. *et al.* A Review of Integrative Imputation for Multi-Omics Datasets. *Front. Genet.* **11**, 570255 (2020).

120. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet. 2019 514* **51**, 592–599 (2019).

121. Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nat. 2023 6207972* **620**, 47–60 (2023).

122. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods 2014 113* **11**, 333–337 (2014).

123. Nguyen, T., Tagett, R., Diaz, D. & Draghici, S. A novel approach for data integration and disease subtyping. *Genome Res.* **27**, 2025–2039 (2017).

124. Röder, B., Kersten, N., Herr, M., Speicher, N. K. & Pfeifer, N. web-rMKL: a web server for dimensionality reduction and sample clustering of multi-view data based on unsupervised multiple kernel learning. *Nucleic Acids Res.* **47**, W605–W609 (2019).

125. Mariette, J. & Villa-Vialaneix, N. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics* **34**, 1009–1015 (2018).

126. Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *https://doi.org/10.1214/12-AOAS597* **7**, 523–542 (2013).

127. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, 8124 (2018).

128. Wu, D., Wang, D., Zhang, M. Q. & Gu, J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics 2015 161* **16**, 1–10 (2015).

129. Ozturk, K., Dow, M., Carlin, D. E., Bejar, R. & Carter, H. The Emerging Potential for Network Analysis to Inform Precision Cancer Medicine. *J. Mol. Biol.* **430**, 2875–2899 (2018).

130. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. in *Bioinformatics* vol. 18 (Oxford University Press, 2002).

131. Paci, P. *et al.* Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. *npj Syst. Biol. Appl.* **7**, 1–11 (2021).

132. Badkas, A., De Landtsheer, S. & Sauter, T. Construction and contextualization approaches for protein-protein interaction networks. *Comput. Struct. Biotechnol. J.* **20**, 3280–3290 (2022).

133. Garcia-Garcia, J., Guney, E., Aragues, R., Planas-Iglesias, J. & Oliva, B. Biana: A software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics* **11**, 1–12 (2010).

134. Mirela-Bota, P. *et al.* Galaxy InteractoMIX: An Integrated Computational Platform for the Study of Protein–Protein Interaction Data. *J. Mol. Biol.* **433**, 166656 (2021).

135. Li, A. & Horvath, S. Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics* **23**, 222–231 (2007).

136. Fox, A. D., Hescott, B. J., Blumer, A. C. & Slonim, D. K. Connectedness of PPI network neighborhoods identifies regulatory hub proteins. *Bioinformatics* **27**, 1135–1142 (2011).

137. Ritz, A. *et al.* Pathways on demand: Automated reconstruction of human signaling networks. *npj Syst. Biol. Appl.* **2**, 1–9 (2016).

138. Lachmann, A. & Ma'ayan, A. Lists2Networks: Integrated analysis of gene/protein lists. *BMC Bioinformatics* **11**, 1–9 (2010).

139. Blokh, D., Segev, D. & Sharan, R. The Approximability of Shortest Path-Based Graph Orientations of Protein–Protein Interaction Networks. *https://home.liebertpub.com/cmb* **20**, 945–957 (2013).

140. Tuncbag, N. *et al.* Network Modeling Identifies Patient-specific Pathways in Glioblastoma. *Sci. Rep.* **6**, 1–12 (2016).

141. Dincer, C., Kaya, T., Keskin, O., Gursoy, A. & Tuncbag, N. 3D spatial organization and network-guided comparison of mutation profiles in Glioblastoma reveals similarities across patients. *PLoS Comput. Biol.* **15**, (2019).

142. Unsal-Beyge, S. & Tuncbag, N. Functional stratification of cancer drugs through integrated network similarity. *npj Syst. Biol. Appl. 2022 81* **8**, 1–13 (2022).

143. Hu, G., Wu, Z., Uversky, V. N. & Kurgan, L. Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions. *Int. J. Mol. Sci. 2017, Vol. 18, Page 2761* **18**, 2761 (2017).

144. McGillivray, P. *et al.* Network Analysis as a Grand Unifier in Biomedical Data Science. *https://doi.org/10.1146/annurev-biodatasci-080917-013444* **1**, 153–180 (2018).

145. Chicco, D. & Jurman, G. A brief survey of tools for genomic regions enrichment analysis. *Front. Bioinforma.* **2**, 968327 (2022).

146. Allman, A., Tang, W. & Daoutidis, P. *Towards a Generic Algorithm for Identifying High-Quality Decompositions of Optimization Problems*. *Computer Aided Chemical Engineering* vol. 44 (Elsevier Masson SAS, 2018).

147. Guimerà, R. & Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nat. 2005 4337028* **433**, 895–900 (2005).

148. Holme, P., Huss, M. & Jeong, H. Subnetwork hierarchies of biochemical pathways. *Bioinformatics* **19**, 532–538 (2003).

149. Zhang, X. S. *et al.* Modularity optimization in community detection of complex networks. *Europhys. Lett.* **87**, 38002 (2009).

150. Redekar, S. S. & Varma, S. L. A Survey on Community Detection Methods and its Application in Biological Network. *Proc. - Int. Conf. Appl. Artif. Intell. Comput. ICAAIC 2022* 1030–1037 (2022) doi:10.1109/ICAAIC53929.2022.9792913.

151. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

152. Yao, B., Zhu, J., Ma, P., Gao, K. & Ren, X. A Constrained Louvain Algorithm with a Novel Modularity. *Appl. Sci. 2023, Vol. 13, Page 4045* **13**, 4045 (2023).

153. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Reports 2019 91* **9**, 1–12 (2019).

154. Aldecoa, R. & Marín, I. Deciphering Network Community Structure by Surprise. *PLoS One* **6**, e24195 (2011).

155. Maleki, F., Ovens, K., Hogan, D. J. & Kusalik, A. J. Gene Set Analysis: Challenges, Opportunities, and Future Research. *Front. Genet.* **11**, 531777 (2020).

156. Katz, S. *et al.* SIGNAL: A web-based iterative analysis platform integrating pathway and network approaches optimizes hit selection from genome-scale assays. *Cell Syst.* **12**, 338-352.e5 (2021).

157. Yang, L., Liu, J., Lu, Q., Riggs, A. D. & Wu, X. SAIC: An iterative clustering approach for analysis of single cell RNA-seq data. *BMC Genomics* **18**, 9–17 (2017).

158. Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* **28**, i451–i457 (2012).

159. Maciejewski, H. Gene set analysis methods: statistical models and methodological differences. *Brief. Bioinform.* **15**, 504–518 (2014).

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

161. Milenković, T. & Pržulj, N. Uncovering Biological Network Function via Graphlet Degree Signatures: *Cancer Inform.* **6**, 257–273 (2008).

162. Pržulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**, e177–e183 (2007).

163. Martin, A. J. M., Dominguez, C., Contreras-Riquelme, S., Holmes, D. S. & Perez-Acle, T. Graphlet Based Metrics for the Comparison of Gene Regulatory Networks. *PLoS One* **11**, e0163497 (2016).

164. Zhu, D. & Qin, Z. S. Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics* **6**, 1–12 (2005).

165. Mangan, S. & Alon, U. Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 11980–11985 (2003).

166. Jazayeri, A. & Yang, C. C. Motif discovery algorithms in static and temporal networks : A survey. 1–38 (2020) doi:10.1093/comnet/cnaa031.

167. Aparício, D., Ribeiro, P. & Silva, F. Graphlet-orbit Transitions (GOT): A fingerprint for temporal network comparison. *PLoS One* **13**, 1–24 (2018).

168. Pržulj, N., Corneil, D. G. & Jurisica, I. Modeling interactome: scale-free or geometric? *Bioinformatics* **20**, 3508–3515 (2004).

169. Levi, H., Elkon, R. & Shamir, R. DOMINO: a network-based active module identification algorithm with reduced rate of false calls. *Mol. Syst. Biol.* **17**, 1–16 (2021).

170. Fredrickson, M. M. & Chen, Y. Permutation and randomization tests for network analysis. *Soc. Networks* **59**, 171–183 (2019).

171. Grasso, R., Micale, G., Ferro, A. & Pulvirenti, A. MODIT: MOtif DIscovery in Temporal Networks. *Front. Big Data* **4**, 806014 (2022).

172. Kim, M. S., Kim, J. R., Kim, D., Lander, A. D. & Cho, K. H. Spatiotemporal

network motif reveals the biological traits of developmental gene regulatory networks in Drosophila melanogaster. *BMC Syst. Biol.* **6**, 1–10 (2012).

173. Rubel, T. & Ritz, A. Augmenting Signaling Pathway Re-constructions. in *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* vol. 10 1–10 (ACM, 2020).

174. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: Integration, analysis and exploration of pathway data. *Nucleic Acids Res.* **48**, D489–D497 (2020).

175. Kamburov, A. *et al.* ConsensusPathDB: Toward a more complete picture of cell biology. *Nucleic Acids Res.* **39**, D712 (2011).

176. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).

177. Ceccarelli, F. *et al.* Bringing data from curated pathway resources to Cytoscape with OmniPath. *Bioinformatics* **36**, 2632–2633 (2020).

178. Kuzmin, K., Gaiteri, C. & Szymanski, B. K. Synergy landscapes: A multilayer network for collaboration in biological research. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 9564 205–212 (Springer Verlag, 2016).

179. Meyer, M. J. *et al.* Interactome INSIDER: a structural interactome browser for genomic studies HHS Public Access. *Nat Methods* **15**, 107–114 (2018).

180. Burley, S. K. *et al.* Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).

181. Mosca, R., Céol, A. & Aloy, P. Interactome3D: Adding structural details to protein networks. *Nat. Methods* **10**, 47–53 (2013).

182. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nature Reviews Cancer* 1–18 (2020) doi:10.1038/s41568-020-0290-x.

183. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

184. Tkačik, G., Walczak, A. M. & Bialek, W. Optimizing information flow in small genetic networks. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **80**, 1–18 (2009).

185. Azpeitia, E., Balanzario, E. P. & Wagner, A. Signaling pathways have an inherent need for noise to acquire information. *BMC Bioinformatics* **21**, (2020).

186. Dou, B. *et al.* Machine Learning Methods for Small Data Challenges in Molecular Science. *Chem. Rev.* **123**, 8736–8780 (2023).

187. Barabási, A. L. Scale-free networks: A decade and beyond. *Science* vol. 325 412–413 (2009).

188. Alm, J. F. & Mack, K. M. L. Degree-correlation, robustness, and vulnerability in finite scale-free networks. *Asian Res. J. Math.* **2**, 1–6 (2016).

189. Page, L. B. S. M. and W. *The PageRank Citation Ranking: Bringing Order to the Web*. (1998).

190. Ghulam, A., Lei, X., Guo, M. & Bian, C. Disease-pathway association prediction based on random walks with restart and pagerank. *IEEE Access* **8**, 72021–72038 (2020).

191. Nitsch, D., Gonçalves, J. P., Ojeda, F., de Moor, B. & Moreau, Y. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics* **11**, 460 (2010).

192. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. in *Journal of Computational Biology* vol. 18 507–522 (J Comput Biol, 2011).

193. Creighton, C. J. *et al.* Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).

194. Tuncbag, N. *et al.* Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. in *Journal of Computational Biology* vol. 20 124–136 (Mary Ann Liebert, Inc., 2013).

195. Kandasamy, K. *et al.* NetPath: A public resource of curated signal transduction pathways. *Genome Biol.* **11**, (2010).

196. Bro, R. & Smilde, A. K. Principal component analysis. *Anal. Methods* **6**, 2812–2831 (2014).

197. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* **6**, 91–97 (2009).

198. von Mering, C. *et al.* STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**, D433–D437 (2005).

199. Schaefer, M. H., Serrano, L. & Andrade-Navarro, M. A. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front. Genet.* **6**, (2015).

200. Porras, P. *et al.* Towards a unified open access dataset of molecular interactions. *Nat. Commun.* **11**, 1–12 (2020).

201. Huang, Z. *et al.* Salmon: Survival analysis learning with multi-omics neural networks on breast cancer. *Front. Genet.* **10**, 166 (2019).

202. Koh, H. W. L. *et al.* iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. *npj Syst. Biol. Appl.* **5**, 22 (2019).

203. Waks, Z. *et al.* Driver gene classification reveals a substantial overrepresentation of tumor suppressors among very large chromatin-regulating proteins. *Sci. Rep.* **6**, 1–12 (2016).

204. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).

205. Zsákai, L. *et al.* Targeted drug combination therapy design based on driver genes. *Oncotarget* **10**, 5255–5266 (2019).

206. Schmidt, T., Bergner, A. & Schwede, T. Modelling three-dimensional protein structures for applications in drug design. *Drug Discovery Today* vol. 19 890–897 (2014).

207. Nero, T. L., Parker, M. W. & Morton, C. J. Protein structure and computational drug discovery. *Biochemical Society Transactions* vol. 46 1367–1379 (2018).

208. Hicks, M., Bartha, I., Di Iulio, J., Craig Venter, J. & Telenti, A. Functional characterization of 3D protein structures informed by human genetic diversity. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8960–8965 (2019).

209. Berman, H. M. *et al. The Protein Data Bank. Nucleic Acids Research* vol. 28 http://www.rcsb.org/pdb/status.html (2000).

210. Venko, K., Roy Choudhury, A. & Novič, M. Computational Approaches for Revealing the Structure of Membrane Transporters: Case Study on Bilitranslocase. *Computational and Structural Biotechnology Journal* vol. 15 232–242 (2017).

211. Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome Networks and Human Disease. *Cell* **144**, 986–998 (2011).

212. Sevimoglu, T. & Yalcin Arga, K. The role of protein interaction networks in systems biomedicine. *CSBJ* **11**, 22–27 (2014).

213. Boughorbel, S., Jarray, F. & El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One* **12**, e0177678 (2017).

214. Magnano, C. S. & Gitter, A. Automating parameter selection to avoid implausible biological pathway models. *npj Syst. Biol. Appl.* **7**, 1–12 (2021).

215. Ricotta, C., Podani, J. & Pavoine, S. A family of functional dissimilarity measures for presence and absence data. *Ecol. Evol.* **6**, 5383–5389 (2016).

216. Sjölund, J., Manetopoulos, C., Stockhausen, M. T. & Axelson, H. The Notch pathway in cancer: Differentiation gone awry. *Eur. J. Cancer* **41**, 2620–2629 (2005).

217. Bazzoni, R. & Bentivegna, A. Role of Notch Signaling Pathway in

Glioblastoma Multiforme Pathogenesis. *Cancers (Basel).* **11**, 292 (2019).

218. Guo, J., Li, P., Liu, X. & Li, Y. NOTCH signaling pathway and non-coding RNAs in cancer. *Pathology Research and Practice* vol. 215 152620 (2019).

219. Qiu, H. *et al.* Notch1 Autoactivation via Transcriptional Regulation of Furin, Which Sustains Notch1 Signaling by Processing Notch1-Activating Proteases ADAM10 and Membrane Type 1 Matrix Metalloproteinase. *Mol. Cell. Biol.* **35**, 3622–3632 (2015).

220. Ramadan, E., Perincheri, S. & Tuck, D. Crosstalk measures for analyzing biological networks in breast cancer. *2010 ACM Int. Conf. Bioinforma. Comput. Biol. ACM-BCB 2010* 579–586 (2010) doi:10.1145/1854776.1854885.

221. Rawlings, J. S., Rosler, K. M. & Harrison, D. A. The JAK/STAT signaling pathway. *J. Cell Sci.* **117**, 1281–1283 (2004).

222. Liu, W., Singh, S. R. & Hou, S. X. JAK-STAT is restrained by Notch to control cell proliferation of the drosophila intestinal stem cells. *J. Cell. Biochem.* **109**, 992–999 (2010).

223. Hilafu, H., Safo, S. E. & Haine, L. Sparse reduced-rank regression for integrating omics data. *BMC Bioinformatics* **21**, (2020).

224. Flores, J. E. *et al.* Missing data in multi-omics integration: Recent advances through artificial intelligence. *Front. Artif. Intell.* **6**, 1098308 (2023).

225. Liu, A. *et al.* From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *npj Syst. Biol. Appl.* **5**, 1–10 (2019).

226. Aragues, R., Sali, A., Bonet, J., Marti-Renom, M. A. & Oliva, B. Characterization of Protein Hubs by Inferring Interacting Motifs from Protein Interactions. *PLOS Comput. Biol.* **3**, e178 (2007).

227. Akavia, U. D. *et al.* An Integrated Approach to Uncover Drivers of Cancer. *Cell* **143**, 1005–1017 (2010).

228. Boyle, E. I. *et al.* GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715 (2004).

229. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* **6**, 271-281.e7 (2018).

230. Penrose, M. D. Connectivity of soft random geometric graphs. *Ann. Appl. Probab.* **26**, 986–1028 (2016).

231. Magnano, C. S. & Gitter, A. Automating parameter selection to avoid implausible biological pathway models. *bioRxiv* 845834 (2019) doi:10.1101/845834.

232. Panni, S., Lovering, R. C., Porras, P. & Orchard, S. Non-coding RNA

regulatory networks. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1863**, 194417 (2020).

233. Khanin, R. & Wit, E. How scale-free are biological networks. *J. Comput. Biol.* **13**, 810–818 (2006).

234. Albert, R., Jeong, H. & Barabási, A. L. Error and attack tolerance of complex networks. *Nat. 2000 4066794* **406**, 378–382 (2000).

235. Fu, D., Hu, Z., Xu, X., Dai, X. & Liu, Z. Key signal transduction pathways and crosstalk in cancer: Biological and therapeutic opportunities. *Transl. Oncol.* **26**, 101510 (2022).

236. Hon, K. W., Zainal Abidin, S. A., Othman, I. & Naidu, R. The Crosstalk Between Signaling Pathways and Cancer Metabolism in Colorectal Cancer. *Front. Pharmacol.* **12**, 768861 (2021).

237. Liu, C. H. *et al.* Analysis of protein-protein interactions in cross-talk pathways reveals CRKL protein as a novel prognostic marker in hepatocellular carcinoma. *Mol. Cell. Proteomics* **12**, 1335–1349 (2013).

238. Charmpi, K., Chokkalingam, M., Johnen, R. & Beyer, A. Optimizing network propagation for multi-omics data integration. *PLOS Comput. Biol.* **17**, e1009161 (2021).

239. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* vol. 18 551–562 (2017).

240. Hu, J. X., Thomas, C. E. & Brunak, S. Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet. 2016 1710* **17**, 615–629 (2016).

241. Ietswaart, R., Gyori, B. M., Bachman, J. A., Sorger, P. K. & Churchman, L. S. GeneWalk identifies relevant gene functions for a biological context using network representation learning. *Genome Biol.* **22**, 55 (2021).

242. Fisch, K. M. *et al.* Omics Pipe: a community-based framework for reproducible multi-omics data analysis. *Bioinformatics* **31**, 1724–1728 (2015).

243. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016).

244. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **8**, 289 (2016).

245. Cieslak, M. C., Castelfranco, A. M., Roncalli, V., Lenz, P. H. & Hartline, D. K. t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis. *Mar. Genomics* **51**, 100723 (2020).

246. Zhou, Y. *et al.* TTD: Therapeutic Target Database describing target

druggability information. *Nucleic Acids Res.* (2023) doi:10.1093/NAR/GKAD751.

247. Stephenson, J. D., Laskowski, R. A., Nightingale, A., Hurles, M. E. & Thornton, J. M. VarMap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. *Bioinformatics* **35**, 4854–4856 (2019).

248. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 1–8 (2018).

249. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).

250. Milacic, M. *et al.* The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res.* **2023**, 1–7 (2023).

251. Forés-Martos, J. *et al.* Transcriptomic metaanalyses of autistic brains reveals shared gene expression and biological pathway abnormalities with cancer. *Mol. Autism* **10**, 1–16 (2019).

252. Levine, D. M. *et al.* Pathway and gene-set activation measurement from mRNA expression data: The tissue distribution of human pathways. *Genome Biol.* **7**, 1–17 (2006).

253. Rydenfelt, M., Klinger, B., Klünemann, M. & Blüthgen, N. SPEED2: inferring upstream pathway activity from differential gene expression. *Nucleic Acids Res.* **48**, W307–W312 (2020).

254. Ke, X. *et al.* Individualized pathway activity algorithm identifies oncogenic pathways in pan-cancer analysis. *eBioMedicine* **79**, (2022).

255. Cruz, L., Soares, P. & Correia, M. Ubiquitin-Specific Proteases: Players in Cancer Cellular Processes. *Pharm. 2021, Vol. 14, Page 848* **14**, 848 (2021).

256. Lens, S. M. A. & Medema, R. H. Cytokinesis defects and cancer. *Nat. Rev. Cancer 2018 191* **19**, 32–45 (2018).

257. Li, J., Dallmayer, M., Kirchner, T., Musa, J. & Grünewald, T. G. P. PRC1: Linking Cytokinesis, Chromosomal Instability, and Cancer Evolution. *Trends in Cancer* **4**, 59–73 (2018).

258. Ben-Zion Berliner, M. *et al.* Central nervous system metastases in breast cancer: the impact of age on patterns of development and outcome. *Breast Cancer Res. Treat.* **185**, 423–432 (2021).

259. Haga, R. B. & Ridley, A. J. Rho GTPases: Regulation and roles in cancer cell biology. *Small GTPases* **7**, 207–221 (2016).

260. Hallin, J. *et al.* The KRASG12C inhibitor MRTX849 provides insight toward therapeutic susceptibility of KRAS-mutant cancers in mouse models and patients.

Cancer Discov. **10**, 54–71 (2020).

261. Kim, H. J., Lee, H. N., Jeong, M. S. & Jang, S. B. Oncogenic KRAS: Signaling and Drug Resistance. *Cancers 2021, Vol. 13, Page 5599* **13**, 5599 (2021).

262. Zhang, S. S. & Nagasaka, M. Spotlight on sotorasib (Amg 510) for krasg12c positive non-small cell lung cancer. *Lung Cancer Targets Ther.* **12**, 115–122 (2021).

263. García-Gutiérrez, L. *et al.* Myc stimulates cell cycle progression through the activation of Cdk1 and phosphorylation of p27. *Sci. Reports 2019 91* **9**, 1–17 (2019).

264. Garces de Los Fayos Alonso, I. *et al.* The Role of Activator Protein-1 (AP-1) Family Members in CD30-Positive Lymphomas. *Cancers 2018, Vol. 10, Page 93* **10**, 93 (2018).

265. Degregori, J., Leone, G., Miron, A., Jakoi, L. & Nevins, J. R. Distinct roles for E2F proteins in cell growth control and apoptosis. *Proc. Natl. Acad. Sci.* **94**, 7245–7250 (1997).

266. Tadesse, S., Caldon, E. C., Tilley, W. & Wang, S. Cyclin-Dependent Kinase 2 Inhibitors in Cancer Therapy: An Update. *J. Med. Chem.* **62**, 4233–4251 (2019).

267. Mendoza, M. C., Er, E. E. & Blenis, J. The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. *Trends Biochem. Sci.* **36**, 320–328 (2011).

268. Demeter, M., Derényi, I. & Szöllősi, G. J. Trade-off between reducing mutational accumulation and increasing commitment to differentiation determines tissue organization. *Nat. Commun.* **13**, (2022).

269. Parenti, I., Rabaneda, L. G., Schoen, H. & Novarino, G. Neurodevelopmental Disorders: From Genetics to Functional Pathways. (2020) doi:10.1016/j.tins.2020.05.004.

270. Rauen, K. A. The RASopathies. *https://doi.org/10.1146/annurev-genom-091212-153523* **14**, 355–369 (2013).

271. Yousef, E. M. *et al.* MCM2: An alternative to Ki-67 for measuring breast cancer cell proliferation. *Mod. Pathol. 2017 305* **30**, 682–697 (2017).

272. Wu, W., Wang, X., Shan, C., Li, Y. & Li, F. Minichromosome maintenance protein 2 correlates with the malignant status and regulates proliferation and cell cycle in lung squamous cell carcinoma. *Onco. Targets. Ther.* **11**, 5025–5034 (2018).

273. Mariani, O. *et al.* JUN Oncogene Amplification and Overexpression Block Adipocytic Differentiation in Highly Aggressive Sarcomas. *Cancer Cell* **11**, 361–374 (2007).

274. Yang, J. & Jiang, W. The Role of SMAD2/3 in Human Embryonic Stem Cells. *Front. Cell Dev. Biol.* **8**, 558892 (2020).

275. Hou, Z. *et al.* KLF2 regulates osteoblast differentiation by targeting of

Runx2. *Lab. Investig. 2018 992* **99**, 271–280 (2018).

276. Yang, J. *et al.* Smad4 is required for the development of cardiac and skeletal muscle in zebrafish. *Differentiation* **92**, 161–168 (2016).

277. Nussinov, R. *et al.* Neurodevelopmental disorders , like cancer , are connected to impaired chromatin remodelers , PI3K / mTOR , and PAK1 - regulated MAPK. *Biophys. Rev.* (2023) doi:10.1007/s12551-023-01054-9.

278. Arici, M. K. & Tuncbag, N. Unveiling Hidden Connections in Omics Data via pyPARAGON : an Integrative Hybrid Approach for Disease Network Construction. *biRxiv* (2023) doi:10.1101/2023.07.13.547583.

279. Janjić, V. & Pržulj, N. The Topology of the Growing Human Interactome Data. *J. Integr. Bioinform.* **11**, 27–42 (2017).

280. Liu, G., Wang, H., Chu, H., Yu, J. & Zhou, X. Functional diversity of topological modules in human protein-protein interaction networks. *Sci. Rep.* **7**, (2017).

281. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One* **10**, e0118432 (2015).

282. Atias, N. & Sharan, R. An algorithmic framework for predicting side effects of drugs. in *Journal of Computational Biology* vol. 18 207–218 (J Comput Biol, 2011).

283. Grimes, T., Potter, S. S. & Datta, S. Integrating gene regulatory pathways into differential network analysis of gene expression data. *Sci. Rep.* **9**, (2019).

284. Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. **29**, 199–209 (2013).

285. Lei, C. & Ruan, J. A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity. *Bioinformatics* **29**, 355–364 (2013).

286. Hulovatyy, Y., Solava, R. W. & Milenković, T. Revealing missing parts of the interactome via link prediction. *PLoS One* **9**, e90073 (2014).

287. Alkan, F. & Erten, C. RedNemo: topology-based PPI network reconstruction via repeated diffusion with neighborhood modifications. *Bioinformatics* **33**, 537–544 (2017).

288. Singh, R., Xu, J. & Berger, B. Struct2Net: Integrating structure into protein-protein interaction prediction. in *Proceedings of the Pacific Symposium on Biocomputing 2006, PSB 2006* 403–414 (2006). doi:10.1142/9789812701626_0037.

289. Segura, J., Sorzano, C. O. S., Cuenca-Alba, J., Aloy, P. & Carazo, J. M.

Using neighborhood cohesiveness to infer interactions between protein domains. *Bioinformatics* **31**, 2545–2552 (2015).

290. Yerneni, S., Khan, I. K., Wei, Q. & Kihara, D. IAS: Interaction Specific GO Term Associations for Predicting Protein-Protein Interaction Networks. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **15**, 1247–1258 (2018).

291. Alanis-Lobato, G., Mier, P. & Andrade-Navarro, M. The latent geometry of the human protein interaction network. *Bioinformatics* **34**, 2826–2834 (2018).

292. Madar, A., Greenfield, A., Ostrer, H., Vanden-Eijnden, E. & Bonneau, R. The inferelator 2.0: A scalable framework for reconstruction of dynamic regulatory network models. in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009* 5448–5451 (IEEE Computer Society, 2009). doi:10.1109/IEMBS.2009.5334018.

293. Fontaine, J. F., Priller, F., Barbosa-Silva, A. & Andrade-Navarro, M. A. Génie: Literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res.* **39**, W455–W461 (2011).

294. Kirkley, A., Cantwell, G. T. & Newman, M. E. J. Belief propagation for networks with loops. *Sci. Adv.* **7**, eabf1211 (2021).

295. Kang, Y. *et al.* HN-PPISP: a hybrid network based on MLP-Mixer for protein–protein interaction site prediction. *Brief. Bioinform.* **24**, (2023).

296. Cheng, F. *et al.* Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nat. Genet. 2021 533* **53**, 342–353 (2021).

297. Mosca, E. *et al.* Characterization and comparison of gene-centered human interactomes. *Brief. Bioinform.* **22**, 1–16 (2021).

298. Yaveroäą Lu, Ö. N. *et al.* Revealing the hidden Language of complex networks. *Sci. Rep.* **4**, 1–9 (2014).

299. Sarajlić, A., Malod-Dognin, N., Yaveroğlu, Ö. N. & Pržulj, N. Graphlet-based Characterization of Directed Networks. *Sci. Rep.* **6**, 1–14 (2016).

300. Zhang, L., Liu, T., Chen, H., Zhao, Q. & Liu, H. Predicting lncRNA–miRNA interactions based on interactome network and graphlet interaction. *Genomics* **113**, 874–880 (2021).

301. Peng, J., Zhou, Y. & Wang, K. Multiplex gene and phenotype network to characterize shared genetic pathways of epilepsy and autism. *Sci. Reports 2021 111* **11**, 1–16 (2021).

302. Riller, Q. & Rieux-Laucat, F. RASopathies: From germline mutations to somatic and multigenic diseases. *Biomed. J.* **44**, 422–432 (2021).

303. Ersahin, T., Tuncbag, N. & Cetin-Atalay, R. The PI3K/AKT/mTOR interactive pathway. *Mol. Biosyst.* **11**, 1946–1954 (2015).

304. Vanhaesebroeck, B., Guillermet-Guibert, J., Graupera, M. & Bilanges, B. The emerging mechanisms of isoform-specific PI3K signalling. *Nat. Rev. Mol. Cell Biol. 2010 115* **11**, 329–341 (2010).

305. Thorpe, L. M., Yuzugullu, H. & Zhao, J. J. PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. *Nat. Rev. Cancer 2015 151* **15**, 7–24 (2014).

306. Bi, L., Okabe, I., Bernard, D. J., Wynshaw-Boris, A. & Nussbaum, R. L. Proliferative defect and embryonic lethality in mice homozygous for a deletion in the p110α subunit of phosphoinositide 3-kinase. *J. Biol. Chem.* **274**, 10963–10968 (1999).

307. Peng, X. ding *et al.* Dwarfism, impaired skin development, skeletal muscle atrophy, delayed bone development, and impeded adipogenesis in mice lacking Akt1 and Akt2. *Genes Dev.* **17**, 1352–1365 (2003).

308. Yu, J. S. L. & Cui, W. Proliferation, survival and metabolism: The role of PI3K/AKT/ mTOR signalling in pluripotency and cell fate determination. *Dev.* **143**, 3050–3060 (2016).

309. Murakami, M. *et al.* mTOR Is Essential for Growth and Proliferation in Early Mouse Embryos and Embryonic Stem Cells. *Mol. Cell. Biol.* **24**, 6710–6718 (2004).

310. Eastman, A. E. *et al.* Resolving Cell Cycle Speed in One Snapshot with a Live-Cell Fluorescent Reporter. *Cell Rep.* **31**, (2020).

APPENDICES

APPENDIX A

Cancer signaling pathways in NetPath

	Pathway Name	Node Size	Edge Size
1	Alpha6Beta4Integrin	66	116
2	Androgen Receptor	165	251
3	BCR	137	261
4	BDNF	72	76
5	CRH	24	27
6	EGFR1	231	756
7	FSH	19	18
8	Hedgehog	36	64
9	ID	27	51
10	IL1	43	93
11	IL11	16	22
12	IL2	67	139
13	IL3	70	97
14	IL4	57	91

	Pathway Name	Node Size	Edge Size
15	IL5	30	36
16	IL6	53	83
17	IL7	18	28
18	IL9	13	15
19	KitReceptor	76	109
20	Leptin	55	74
21	Notch	74	154
22	OncostatinM	37	42
23	Prolactin	68	103
24	RAGE	23	25
25	RANKL	57	76
26	TSH	48	47
27	TSLP	7	7
28	TWEAK	17	15
29	TCR	154	271
30	TGFbetaReceptor	209	452
31	TNFalpha	239	473
32	Wnt	106	220

APPENDIX B

Frequently seen biological processes in clusters

GO ID	The number of patients	Name	Cluster
GO:0006511	21	ubiquitin-dependent protein catabolic process	Cluster1
GO:0006605	20	protein targeting	Cluster1
GO:0051056	16	regulation of small GTPase mediated signal transduction	Cluster1
GO:0031146	16	SCF-dependent proteasomal ubiquitin-dependent protein catabolic process	Cluster1
GO:0015031	15	protein transport	Cluster1
GO:0010628	15	positive regulation of gene expression	Cluster1
GO:0008150	13	biological_process	Cluster1
GO:0042659	13	regulation of cell fate specification	Cluster1
GO:0000281	13	mitotic cytokinesis	Cluster1
GO:2000736	13	regulation of stem cell differentiation	Cluster1
GO:0030154	12	cell differentiation	Cluster1
GO:0045087	12	innate immune response	Cluster1
GO:0007399	12	nervous system development	Cluster1
GO:0070936	12	protein K48-linked ubiquitination	Cluster1
GO:0000165	12	MAPK cascade	Cluster1
GO:0007155	12	cell adhesion	Cluster1
GO:0032956	11	regulation of actin cytoskeleton organization	Cluster1
GO:0009410	11	response to xenobiotic stimulus	Cluster1
GO:0006886	10	intracellular protein transport	Cluster1

GO ID	The number of patients	Name	Cluster
GO:0007010	10	cytoskeleton organization	Cluster1
GO:0000281	13	mitotic cytokinesis	Cluster2
GO:0015031	12	protein transport	Cluster2
GO:0030154	12	cell differentiation	Cluster2
GO:0051123	11	RNA polymerase II pre-initiation complex assembly	Cluster2
GO:0060261	11	positive regulation of transcription initiation by RNA polymerase II	Cluster2
GO:0043123	11	positive regulation of I-kappaB kinase/NF-kappaB signaling	Cluster2
GO:0018105	11	peptidyl-serine phosphorylation	Cluster2
GO:0000165	10	MAPK cascade	Cluster2
GO:0007010	10	cytoskeleton organization	Cluster2
GO:0008150	10	biological_process	Cluster2
GO:0043065	10	positive regulation of apoptotic process	Cluster2
GO:0008285	9	negative regulation of cell population proliferation	Cluster2
GO:0050821	9	protein stabilization	Cluster2
GO:0006511	8	ubiquitin-dependent protein catabolic process	Cluster2
GO:0032968	8	positive regulation of transcription elongation by RNA polymerase II	Cluster2
GO:0032922	8	circadian regulation of gene expression	Cluster2
GO:0007266	7	Rho protein signal transduction	Cluster2
GO:0016055	7	Wnt signaling pathway	Cluster2
GO:0006897	7	endocytosis	Cluster2
GO:0007155	7	cell adhesion	Cluster2
GO:2000045	19	regulation of G1/S transition of mitotic cell cycle	Cluster3

GO ID	The number of patients	Name	Cluster
GO:2000819	19	regulation of nucleotide-excision repair	Cluster3
GO:0030071	19	regulation of mitotic metaphase/anaphase transition	Cluster3
GO:2000781	19	positive regulation of double-strand break repair	Cluster3
GO:0070316	19	regulation of G0 to G1 transition	Cluster3
GO:1902459	14	positive regulation of stem cell population maintenance	Cluster3
GO:0007399	14	nervous system development	Cluster3
GO:0045663	13	positive regulation of myoblast differentiation	Cluster3
GO:0045597	13	positive regulation of cell differentiation	Cluster3
GO:0045582	12	positive regulation of T cell differentiation	Cluster3
GO:0006897	11	endocytosis	Cluster3
GO:0006605	10	protein targeting	Cluster3
GO:0045596	10	negative regulation of cell differentiation	Cluster3
GO:0016055	9	Wnt signaling pathway	Cluster3
GO:0006913	9	nucleocytoplasmic transport	Cluster3
GO:0000165	9	MAPK cascade	Cluster3
GO:0006337	8	nucleosome disassembly	Cluster3
GO:0015031	8	protein transport	Cluster3
GO:0006511	8	ubiquitin-dependent protein catabolic process	Cluster3
GO:0030154	8	cell differentiation	Cluster3
GO:0032956	23	regulation of actin cytoskeleton organization	Cluster4
GO:0007155	19	cell adhesion	Cluster4
GO:0008150	18	biological_process	Cluster4
GO:0030154	18	cell differentiation	Cluster4

GO ID	The number of patients	Name	Cluster
GO:0030865	17	cortical cytoskeleton organization	Cluster4
GO:0000165	16	MAPK cascade	Cluster4
GO:0034063	16	stress granule assembly	Cluster4
GO:0006605	16	protein targeting	Cluster4
GO:0006897	14	endocytosis	Cluster4
GO:0015031	14	protein transport	Cluster4
GO:0006417	13	regulation of translation	Cluster4
GO:0007266	13	Rho protein signal transduction	Cluster4
GO:0043065	12	positive regulation of apoptotic process	Cluster4
GO:0051496	12	positive regulation of stress fiber assembly	Cluster4
GO:0018107	12	peptidyl-threonine phosphorylation	Cluster4
GO:0046777	12	protein autophosphorylation	Cluster4
GO:0007010	12	cytoskeleton organization	Cluster4
GO:0032968	11	positive regulation of transcription elongation by RNA polymerase II	Cluster4
GO:0051056	11	regulation of small GTPase mediated signal transduction	Cluster4
GO:0008285	11	negative regulation of cell population proliferation	Cluster4

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name	:	Arıcı, Müslüm Kaan
Nationality	:	Turkish (TC)

EDUCATION

Degree	Department	Institution	Year of Graduation
MS	Biotechnology,	METU	2017
BS	Biology,	METU	2014

WORK EXPERIENCE

Year	Place	Enrollment
2022-2023	National Cancer Institute, USA	Guest Researcher
2018-2022	TUBITAK	Scholarship Student
2014-Preesent	Foot And Mouth Diseases Institute	Biologist
2008-2014	Foot And Mouth Diseases Institute	Laboratory Assistant

PUBLICATIONS

(* first author(s), #corresponding author(s))

1. Arici, M. Kaan*, and Nurcan Tuncbag#. “Unveiling Hidden Connections in Omics Data via pyPARAGON: an Integrative Hybrid Approach for Disease Network Construction,” *bioRxiv* https://doi.org/10.1101/2023.07.13.547583.

2. Yavuz, Bengi Ruken*, M. Kaan Arici*, Habibe Cansu Demirel*, Ruth Nussinov#, Chung-Jung Tsai, Hyunbum Jang, and Nurcan Tuncbag#. 2023. “Neurodevelopmental disorders and cancer networks share pathways, but differ in mechanisms, signaling strength, and outcome” *NPJ Genomic. Medicine*. 8, 37. https://doi.org/10.1038/s41525-023-00377-6.

3. Nussinov, Ruth#*, Bengi Ruken Yavuz, M. Kaan Arici, Habibe Cansu Demirel, Mingzhen Zhang, Yonglan Liu, Chung-Jung Tsai, Hyunbum Jang, and Nurcan Tuncbag. 2023. “Neurodevelopmental Disorders, like Cancer, Are Connected to Impaired Chromatin Remodelers, PI3K/mTOR, and PAK1-Regulated MAPK.” *Biophysical Reviews*, April, 1–19. https://doi.org/10.1007/s12551-023-01054-9.

4. Demirel, Habibe Cansu*, M. Kaan Arici*, and Nurcan Tuncbag#. 2022. “Computational Approaches Leveraging Integrated Connections of Multi-Omic Data toward Clinical Applications.” *Molecular Omics* 18 (1): 7–18. https://doi.org/10.1039/D1MO00158B.

5. Arici, M. Kaan*, and Nurcan Tuncbag#. 2021. “Performance Assessment of the Network Reconstruction Approaches on Various Interactomes.” *Frontiers in Molecular Biosciences*, https://doi.org/10.3389/fmolb.2021.666705.

TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences [ ]

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences [ ]

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics [ ]

Enformatik Enstitüsü / Graduate School of Informatics [X]

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences [ ]

YAZARIN / AUTHOR

Soyadı / Surname : ARICI
Adı / Name : Müslüm Kaan
Bölümü / Department : Tıp Bilişimi /

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English) : Uncovering Hidden Connections and Functional Modules via pyPARAGON: a Hybrid Approach for Network Contextualization

TEZİN TÜRÜ / DEGREE: **Yüksek Lisans** / Master [ ] **Doktora** / PhD [X]

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. [X]
2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two year. *** [ ]
3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. * [ ]

***** *Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir.*
A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature ..... **Tarih** / Date : 22.01.2024