NOVEL USES OF CORRELATION NETWORKS AND CONSENSUS NODE-COMMUNITIES FOR BIOMEDICAL DATA ANALYSIS

Reddy Rani Vangimalla

Doctor of Philosophy Thesis

August 2021

International Institute of Information Technology, Bangalore

NOVEL USES OF CORRELATION NETWORKS AND CONSENSUS NODE-COMMUNITIES FOR BIOMEDICAL DATA ANALYSIS

Submitted to International Institute of Information Technology,
Bangalore
in Partial Fulfillment of
the Requirements for the Award of
Doctor of Philosophy

by

Reddy Rani Vangimalla
PH2016009

International Institute of Information Technology, Bangalore
August 2021

Thesis Certificate

This is to certify that the thesis titled **Novel Uses of Correlation Networks and Consensus Node-Communities for Biomedical Data Analysis** submitted to the International Institute of Information Technology, Bangalore, for the award of the degree of **Doctor of Philosophy** is a bona fide record of the research work done by **Reddy Rani Vangimalla**, **PH2016009**, under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Jaya

Prof. Jaya Sreevalsan Nair

Prof. T. K. Srikanth

Bengaluru,

The 30th of August, 2021.

NOVEL USES OF CORRELATION NETWORKS AND CONSENSUS NODE-COMMUNITIES FOR BIOMEDICAL DATA ANALYSIS

Abstract

The correlation between random variables has been routinely used for analyzing biomedical datasets in life sciences research. These studies are significant for finding biological patterns, disease prognosis, and treatment. We explore novel uses of correlation networks and consensus methods for identifying the node-communities to improve the accuracy of data mining problems, such as finding communities/modules with maximized modularity, retrieving feature-rich subspace of data, identifying tumour subtypes in patients, etc. We study the application of consensus methods in correlation networks in two different biomedical data problems. In the *first study*, using a brain functional connectivity network (FCN), which is a correlation network, the modularly organized brain regions of resting-state functional magnetic resonance imaging (fMRI) data are obtained. Extracting modular brain regions allows practitioners to study spontaneous brain activity. In the *second problem*, a heterogeneous correlation network model is constructed from multi-omic features of cancer data, and a feature-rich integrative subspace of data is retrieved, which serves as an application for predicting cancer-specific patient subtypes. Finding significant genes and subtypes of the diseases is vital for the early prognosis of the disease, personalized treatment; therefore, the improved survival probability of the patient.

In brain FCN, the correlations are computed among the regions of interest (ROIs) derived from a specific parcellation technique. We perform case studies of FCN of the human brain at resting state, with different sizes/resolutions and parcellation atlases (AAL, Schaefer) for finding the modular organization of the FCN. Identifying

modules of brain FCN using a sparsified correlation matrix and using network-theory procedures is a well-researched approach. However, these procedures include loss of correlation information due to sparsification. In general, there is still no consensus in the research community on finding the right threshold value for edge filtering in FCNs. The novelty of our work lies in using a *complete* (*full*) functional connectivity network, for functional segregation of the network. We perform an extensive analysis of the use of exploratory factor analysis (EFA) for community detection in FCN by exploiting the semantics of the correlation matrix. To effectively use EFA, we implement a novel consensus-based algorithm using a multiscale approach, considering the number of factors n_F as a scale. We use an ensemble of experiments and extensive quantitative analysis and its outcomes to identify the optimal set of scales for efficient node-partitioning. Using the multiscale approach, we transform the correlation network to a 'co-association' network. The transformed network sparsifies the dense (full) network, where edges represent the likeliness of nodes clustering together. The multiscale approach and consensus community detection procedures help identify *modularity maximized communities* and *cliques* within communities, *hierarchical modular organization* of communities, and exhibit *hemispheric symmetry* of nodes in communities. Our results of consensus (node-) communities and cliques have been found to be relevant for the brain activity in its resting state, thus concluding the effectiveness of EFA.

In multi-omics studies of cancer data, integrative analysis of multi-omics data is essential for biomedical applications, as it is required for a comprehensive understanding of biological function. Integrating multi-omics data serves multiple purposes, such as, an integrated data model, dimensionality reduction of omic features, patient clustering, etc. However, there is a gap in combining some of the widely used integrative analyses to build more powerful tools. In this work, we propose a multi-level integration algorithm to identify a representative integrative subspace and use it for cancer subtype prediction. The breast and lung cancer multi-omics data of DNA methylation (genome) and

mRNA expression (transcriptome) features from ‘The Cancer Genome Atlas (TCGA)’ database is used in this study. The choice of cancer phenotypes is to investigate our workflow on one of the most widely studied cancer subtypes, i.e., breast cancer data, and one of the least studied cancer subtypes, i.e., lung cancer data. The three integrative approaches we implement on multi-omics features are, (1) multivariate multiple (linear) regression of the features from a cohort of patients/samples, (2) bipartite graph between different features, and (3) fusion of sample similarity networks across the features. We use a type of multilayer network, called heterogeneous network, as a data model to transition between a network-free (NF) regression model and a network-based (NB) fusion model, which uses correlation networks. Our proposed heterogeneous correlation network model, HCNM, is central to our algorithm for gene-ranking, integrative-subspace identification, and tumor-specific subtypes prediction. The genes of our representative integrative subspace have been *enriched with gene-ontology* and found to exhibit *significant gene-disease association* (GDA) scores. The subspace in genes which is less than 10% of the total gene-set of each genomic feature in both phenotypes is used with NB fusion integrative model to predict sample subtypes. As the identified integrative subspace data of multi-omics is less prone to noise, bias, and outliers, our experiments show that the subtypes in our results *agree with benchmark studies* of breast and lung cancer data, and also exhibit *better classification* between poor and good survival of patient cohorts.

Acknowledgements

My sincere gratitude to my esteemed supervisor *Prof. Jaya Sreevalsan Nair*, for her invaluable timely support, supervision, and encouragement. Her expertise and vision were precious and are highly helpful in formulating and solving the research problems. The journey is both enlightening and comfortable, with her assistance, insightful comments, moral support, and suggestions. I would like to extend my sincere thanks to my co-advisor, Prof. T. K. Srikanth, my doctoral committee members Prof. Balakrishnan Ashok and Prof. Manish Gupta for their comments and guidance.

I would like to offer my special thanks to Prof. Olaf Sporns and Joshua Faskowitz for sharing the datasets, providing domain knowledge, and for their consistent and timely help, suggestions, and comments/feedback. I would also like to express my sincere gratitude to Prof. Kyung-Ah Sohn and Dr. Hyun-hwan Jeong for being patient and providing their expertise comments and suggestions. I thank Dr. Rose Dawn Bharath, Dr. Urvakhsh Meherwan Mehta for working closely with us on the real-time neurobiology problems and introducing us to the radiology lab.

I am grateful to the International Institute of Information Technology, Bangalore (IIITB) fraternity at large for their support in my work throughout my stay at the university. I am also thankful for the funding supported by Visvesvaraya Ph.D. Scheme for Electronics & IT by Ministry of Electronics & Information Technology (MeitY), Government of India. I would like to especially acknowledge both MeitY and IIITB for generously providing financial support to me for attending top-tier conferences, such as IEEE EMBC, ISMB, and Eurographics VCBM, for presenting my work.

I would like to offer my thanks to Prof. Chetan Parikh for arranging Ph.D. chit-chat sessions, which were inspiring and motivating. Also, my heartful thanks to the IIITB administration and faculty for providing a pleasant, friendly, and supportive environment. I am deeply thankful to my colleagues, Shivam, Raghavendra, Yoga Laxmi,

Raksha, Praseeda, Rajesh, Harshitha, Komal, and Surya, for being through the thick and thin of my journey in IIITB.

I would like to thank my parents and parents-in-law, brother and his family, my son *Sai Sree Viswagna* and husband *Naga Ashok Jampani* for their enormous support and patience throughout my journey.

List of Publications

[I] Under Review –**Reddy Rani Vangimalla**, and J. Sreevalsan-Nair, "Communities and Cliques in Functional Brain Network Using Multiscale Consensus Approach."

[II] **Reddy Rani Vangimalla**, and J. Sreevalsan-Nair, " HCNM: Heterogeneous Correlation Network Model for Multi-level Integrative Study of Multi-omics Data for Cancer Subtype Prediction." (Accepted) in the Proceedings of the 43rd International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), November 2021.

[III] V. Sridhar, Jaya Sreevalsan-Nair, Pritesh Rajesh Ghogale, and **Reddy Rani Vangimalla**, *"Sharing and Use of Non-Personal Health Information: Case of the COVID-19 Pandemic"*, Chapter 8, In V. Sridhar (Ed.) *Data Centric Living: Algorithms, Digitization and Regulation.* Routledge, Taylor & Francis Group, June, 2021. DOI: 10.4324/9781003093442, ISBN 9780367536534.

[IV] **Reddy Rani Vangimalla** and Jaya Sreevalsan-Nair. "A Multiscale Consensus Method Using Factor Analysis to Extract Modular Regions in the Functional Brain Network," In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 2824–2828. IEEE, 2020. DOI: 10.1109/EMBC44109.2020.9175622.

[V] Urvakhsh Meherwan Mehta, Darshan Shadakshari, Pulaparambil Vani, Shalini S Naik, V Kiran Raj, **Reddy Rani Vangimalla**, YC Janardhan Reddy, Jaya Sreevalsan-Nair, and Rose Dawn Bharath. "Case report: Obsessive compulsive disorder in posterior cerebellar infarction-illustrating clinical and functional connectivity modulation using mri-informed transcranial magnetic stimulation," Sep, 2020. Wellcome Open Res 2020, 5:189, PMID: 32995558, PMCID: PMC7503177. DOI:10.12688/wellcomeopenres.16183.2.

[VI] **Reddy Rani Vangimalla** and Jaya Sreevalsan-Nair. “Comparing community detection methods in brain functional connectivity networks.” In International Conference on Computational Intelligence, Cyber Security, and Computational Models, pages 3–17. Springer, 2019. DOI:10.1007/978-981-15-9700-8_1.

[VII] **Reddy Rani Vangimalla**, and J. Sreevalsan-Nair, “Construction and Visualization of Diseasome of Lung Diseases Associated with COVID-19 from Co-association Networks of Multi-omics Data,” July, 2020. 28th Annual International Conferences of Intelligent Systems for Molecular Biology, ISMB 2020, NetBio. Poster. DOI:10.7490/f1000research.1118138.1.

[VIII] J. Sreevalsan-Nair, **Reddy Rani Vangimalla**, and Pritesh Rajesh Ghogale, “Influence of COVID-19 Transmission Stages and Demographics on Length of In-Hospital Stay in Singapore for the First 1000 Patients,” July, 2020. 28th Annual International Conferences of Intelligent Systems for Molecular Biology ISMB 2020, COVID-19. Poster. DOI:10.7490/f1000research.1118104.1.

[IX] **Reddy Rani Vangimalla**, and J. Sreevalsan-Nair, “Consensus Methods for Network Analysis of Biomedical Data:Case Studies on Brain Functional Connectivity Network and Gene-GeneAssociation Networks,” February, 2020. In 4th International Conference on Computational Intelligence and Networks CINE 2020. Doctoral Symposium. PDF.

[X] Jaya Sreevalsan-Nair, **Reddy Rani Vangimalla**, and Pritesh Rajesh Ghogale. “Analysis and estimation of length of in-hospital stay using demographic data of covid-19 recovered patients in singapore,” medRxiv, April, 2020. Preprint

[XI] **Reddy Rani Vangimalla**, and J. Sreevalsan-Nair, “RadTrix: A Composite Hybrid Visualization for Unbalanced Bipartite Graphs in Biological Datasets,” 9th Eurographics Workshop on Visual Computing for Biology and Medicine, September 2019. Poster. Conference proceedings.

[XII] Jaya Sreevalsan-Nair, Shivam Agarwal, **Reddy Rani Vangimalla**, Sanat Ramesh,and Nirmala Murthy. “Collaborative design of visual analytic techniques for sur-vey data for community-based research in public health,” 8th Workshop on Visual Analytics in Healthcare, affiliated with IEEE VIS 2017. Poster.

Contents

List of Figures

List of Tables

List of Abbreviations

AAL	Automated Anatomical Labeling
ANF	Affinity Network Fusion
BCT	Brain Connectivity Toolbox
BNV	BrainNet Viewer
BOLD	Blood-Oxygenation Level-Dependent
BRCA	Breast invasive carcinoma
DI	Dunn Index
DNA	Deoxyribonucleic acid
EC	Effective Connectivity
EFA	Exploratory Factor Analysis
EEG	Electroencephalography
FC	Functional Connectivity
FCN	Functional Connectivity Network
FL	Factor Loading
fMRI	Functional magnetic resonance imaging
FRC	Factor Retention Criterion
GDA	Gene-Disease Association
HC	Hierarchical consensus clustering
h-clust	Hierarchical clustering

HCNM	Heterogeneous Correlation Network Model
ICGC	International Cancer Genome Consortium
IIITB	International Institute of Information Technology Bangalore
IM	Infomap Method
KMO	Kaiser-Meyer-Olkin’s
Lasso	Least absolute shrinkage and selection operator
LM	Louvain Method
LUSC	Lung squamous cell carcinoma
MC	Multiscale consensus
MEG	Magnetoencephalography
miRNA	microRNA
MI	Mutual Information
MLE	Maximum Likelihood Estimation
MLN	Multilayer Networks
MMR	Multivariate Multiple (linear) Regression
mRNA	messengerRNA
MSA	Measure of Sampling Adequacy
NB	Network-Based
NF	Network-Free
NMI	Normalized Mutual Information

OCD	Obsessive-Compulsive Disorder
PA	Percolation Analysis
PCAWG	Pan-Cancer Analysis of Whole Genomes
Q	Modularity
RMSEA	Root Mean Square Error of Approximation
RMSR	Root Mean Square Residual
RNA	Ribonucleic acid
ROIs	Regions Of Interest
rs-fMRI	resting-state Functional magnetic resonance imaging
R2E	Rank-two ellipse
S	Silhouette-score
SC	Structural Connectivity
SNF	Similarity Network Fusion
TCGA	The Cancer Genomic Atlas
WGCNA	WeiGhted Correlation Network Analysis

CHAPTER 1

INTRODUCTION

In biomedical research, correlation analysis is used to understand the relationship between the independent variables/risk factors/predictor variables, with dependent variable/disease outcome. Correlation measure between two variables provides statistical evidence of their strength/association. Sir Francis Galton [7], in his work "Co-relations and their measurement, chiefly from Anthropometric Data" had first proposed the notion of correlation, which is later mathematically described by Karl Pearson [8]. The biomedical data problems have been extensively studied for several decades using networks. The network is represented as a graph (G) with vertices (V) and edges (E), where edges signify the strength/association between the vertices. The graph theory was first introduced by Leonhard Euler in 1735 while solving the Königsberg bridge problem [9, 10]. For the last two decades, graph theory along with statistical mechanics, have paved the way to network science, which is applied in complex systems such as cellular networks, neural networks, social networks, trade networks, ecological networks, communication networks, biological networks, etc [11]. In our work, we use correlations between the vertices as edges to study some biomedical problems.

There exist various biomedical problems that consider correlation and network-based analysis. For example, Epigenomics correlation study to learn hyper and hypomethylated CpG islands that help to study reproductomics [12], to learn drug-disease associations using gene expression data [13–15], to understand perinatal and neonatal adversaries using the features such as maternal age and parity [16, 17], to study heart stroke severity using EHR data [18, 19], to learn the associations between lung cancer and cigarette smoking habit [20, 21]. Correlation networks serve to give a broad overview of an organism's state, be it on the metabolic level, the proteomic level, the transcript level, or a combination of these [22].

Researchers often have an idea of what sort of relationship they are looking at, such as a linear relationship or a step-function response to a perturbation in time series data, etc. These relationships are computed using several methods, e.g., Pearson's correlation, mutual information, phase coherence, partial correlation, granger causality, coherence, Spearman's rank correlation, median-based correlation measures, maximal information criteria (MIC), Kendall rank correlation, etc. Most often, in biology, Pearson's correlation measure is used. The normalization/standardization procedure of data prepossessing operation makes the variables normally distributed, making it feasible to use Pearson's correlation coefficient. Pearson's correlation measure expects few prerequisites such as linearity (the straight-line relationship between each of the two variables), normality (variables with a bell-shaped curve distribution), homoscedasticity (the data is equally distributed about the regression line), and outlier verification. Non-normally distributed data may include outlier values that necessitate the usage of Spearman's correlation coefficient [23]. In summary, correlation coefficients are used to assess the strength between the pairs of variables/vertices/nodes.

In biomedical and healthcare scientific study, neurobiological research and cancer research is at the top of the list. In this doctoral work, by examining correlation-based networks, we address some of the open problems in these two domains, a) finding functional segregation of the brain functional connectivity network at resting state and b) identifying the significant subspace of multi-omics cancer data that aid in finding subtypes in cancer. The former is derived from functional magnetic resonance imaging (fMRI) data, and the latter is multi-omics data of subjects of a specific cancer type downloaded from The Cancer Genomic Atlas (TCGA) database[1].

1.1 Motivation

Our motivation in this study is to use correlation networks of biology and improve the accuracy of data mining problems by applying consensus community detection procedure. We study the application of consensus methods in correlation networks, a) in brain Functional Connectivity Networks (FCN), and b) as heterogeneous multi-omics networks of cancer subjects.

In brain FCN, the motivation is to, i) determine modularly organized brain regions (communities), leading to understanding spontaneous connections of resting-state fMRI data, and ii) find functional segregation of the correlation matrix. Functional segregation refers to clique or communities or motifs or the number of triangles of a network, explaining the extent to which a network can form into separate components [24]. Functional segregation is synonymous to a community in the network-science that refers to neuronal processing carried out among communities. The *communities*/modules/node-groupings are defined by dense intra-community connections and sparse inter-community connections. A *clique* is a completely connected network with a tight-knit of nodes, where every node has links between every pair of nodes. This line of study has been around for 2-3 decades [25–27]. The research community is working on fine-tuning the methods for studying complex problems in the presence of brain disorders, such as Alzheimer's, autism, epilepsy, Parkinson's, obsessive compulsory disorder, modular organization changes in the brain with age etc [28–34].

Finding significant cancer-specific genes and subtypes of cancer is vital for the early prognosis of the disease, personalized treatment; therefore, it improves the survival

[1] https://portal.gdc.cancer.gov/

probability of the patient. For the second study, using high dimensional multi-omics cancer data and by employing multi-level integrative algorithms, we aim to find, i) a representative integrative subspace with feature-rich genes that are less prone to biases, noise, and outliers, and ii) subtypes of patients using the subspace of multi-omics data. We investigate our workflow on one of the most widely studied cancer subtypes, i.e., breast cancer, and one of the least studied cancer subtypes, i.e., lung cancer data.

1.2 Contribution

Functional Segregation in Brain FCN: The brain FCN communities are well studied using network theory measures on the sparsified/discretized/thresholded network [35–37]. The threshold value used for edge filtering, if not chosen optimally, may lead to loss of information. Hence, with the sparsified network, the semantics of the correlation network is not utilized thoroughly to infer appropriate knowledge. In general, there is no consensus on the science behind the choice of thresholds [36]. We address this problem by considering the full/complete correlation network for finding modularly organized brain regions. We propose using a weighted, completely connected, and undirected correlation network and applying Exploratory Factor Analysis (EFA) to find the node-partitions in the FCN. Charles Spearman [38] first proposed factor analysis; in the domain of psychology study. EFA works with an assumption that there exists an underlying structure among the variables. If the correlation values between the variables are not significant, then EFA fails to identify groups of variables, owing to the absence of structure among the variables. EFA expects an input parameter, namely, the number of factors, n_F, for its implementation. However, there is no *ground truth* for a definite number of node-partitions in FCN. Hence, we decide not to rely on a single value of n_F for EFA. We define the value of n_F as the *scale* for EFA, and propose to apply EFA for multiple scales. A single-scale EFA, in general, suffers from the issues such as

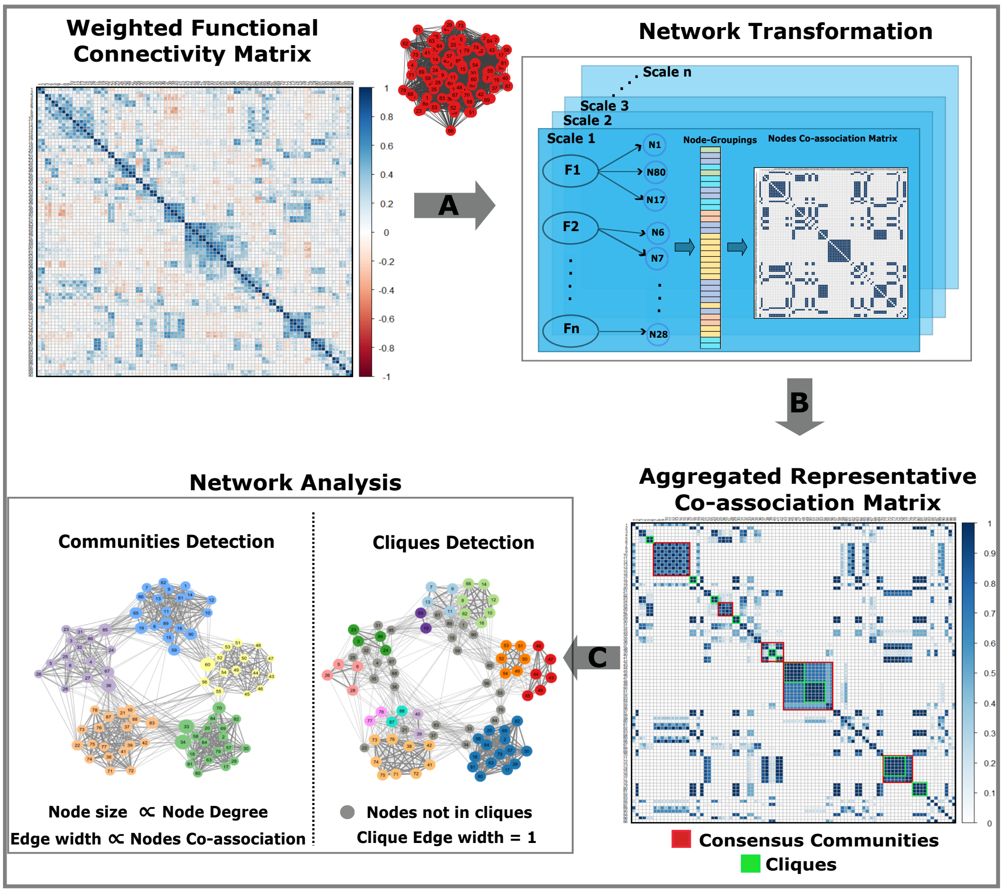


Figure FC1.1: Our proposed workflow of computing node-partitioning using EFA on the correlation matrix of FCN, at multiple scales, transforming the FCN to a multiscale co-association matrix, and performing generalized Louvain method [1] for community detection.

replicability of node-partitioning [39], and the generalizability of the workflow [40]. To address these issues and add to the exploratory and experimental characteristics of EFA, we find a range of n_F and perform EFA with multiple scales, as shown in Figure FC1.1. The node-partitioning is represented using a co-association matrix D^k, for the k^{th} scale. We choose a set of values of n_F to compute EFA and the communities/node-groupings are aggregated to generate a final co-association matrix, D, using consensus voting of the D^k at $k = 1, 2, \ldots, N$ scales. Thus, we transform a weighted fully connected FCN (correlation matrix) to a representative co-association matrix. The D is symmetric, with the values in the range of 0 to 1, 0 for the nodes that did not group together in any of the node-partitionings obtained in any of the chosen scales, and 1 for nodes that always belonged in the same community through all the node-partitionings. Thus, we use D as an adjacency matrix of our *transformed* network [41]. Generalized Louvain (genLouvain) [1] community detection algorithm is then implemented on the transformed network to find the consensus-based communities and cliques of FCN. The proposed workflow of finding functional segregation by identifying communities and cliques of FCN using the multiscale EFA method is depicted in Figure FC1.1

Multi-level Integrative Study of Multi-omics Cancer Data: In multi-omics studies of cancer data, integrative analysis of multi-omics data is essential for biomedical applications, as it is required for a comprehensive understanding of biological function. For cancer studies pertaining to outcome prediction, multi-omics information has been routinely integrated at the data level to obtain transformed data models, such as, regression and network models [42–44]. The available high-throughput omic data causes a "small n, large p" or "short-fat data" problem. This enforces the need to find the representative subspace of multi-omics features [43,45]. Multilayer networks (MLN) are less studied in biology, whereas network analysis of a single layer is widely investigated for protein interactions, metabolic associations, gene co-expressions, pathways in regulatory, etc. In this work, we propose a workflow to address these problems using *multi-level integrative procedures* and *heterogeneous networks* of *multi-omics* cancer data (Figure FC1.2). The heterogeneous networks are a special class of multilayer networks, that consist of intra-layer and inter-layer graphs, the latter being bipartite graphs [46,47]. In order to achieve a multi-level integration of the multi-omics data through existing integrative methods, namely regression model, which is a network-free method (I_1 in Figure FC1.2) and network-based fusion method (I_3 in Figure FC1.2), we propose a data model that will transition one method to another, referred to as the Heterogeneous Correlation Network Model (`HCNM`) (I_2 in Figure FC1.2). We work with DNA methylation and mRNA expression data of TCGA and generate intra-layer graphs that are heterogeneous correlation networks for each omic feature (Figure FC1.2). The commu-

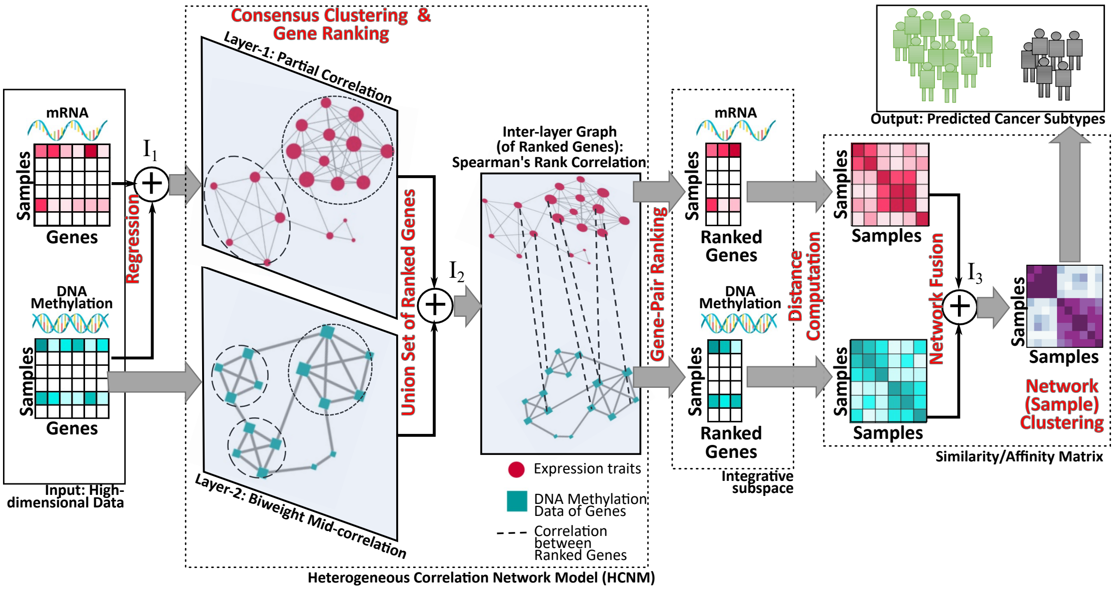

Consensus Clustering &
Gene Ranking
Layer-1: Partial Correlation
Layer-2: Biweight Mid-correlation
mRNA
Samples
Genes
DNA Methylation
Samples
Genes
Input: High-dimensional Data
I1
Regression
Union Set of Ranked Genes
I2
Inter-layer Graph (of Ranked Genes): Spearman's Rank Correlation
● Expression traits
■ DNA Methylation Data of Genes
- - - Correlation between Ranked Genes
Heterogeneous Correlation Network Model (HCNM)
Gene-Pair Ranking
mRNA
Samples
Ranked Genes
DNA Methylation
Samples
Ranked Genes
Integrative subspace
Distance Computation
Samples
Samples
Samples
Samples
Network Fusion
I3
Samples
Samples
Network (Sample) Clustering
Similarity/Affinity Matrix
Output: Predicted Cancer Subtypes

Figure FC1.2: Illustration of our proposed model. We use multi-level integration of multi-omics data using HCNM, with I_1, I_2, and I_3 as the different integration steps.

nities of genes in each intra-layer graph are identified. The genes within a community are known to have high modularity value and, therefore, a high likelihood for similar behavior or influence on diseases. We use three characteristically different community detection procedures and identify the final set of ranked genes using consensus voting of all three models. The ranked genes are used to generate an inter-layer cross-correlations graph. The representative integrative subspace of multi-omics is derived by ranking the gene-pairs of the inter-layer graph. The patient subtypes in cancer are studied with the final identified feature-rich integrative subspace and by applying network-based fusion integrative procedures. Overall, we find representative integrative subspace, and cancer subtypes, by multi-level integration of multi-omics data. We imply three occurrences of multi-omics integration in our algorithm: I_1 when using a multivariate multiple (linear) regression (MMR) model, I_2 for selecting genes to compute the inter-layer graph using cross-correlations, and I_3 for network fusion of similarity networks of samples/patients. We propose a three-level integration algorithm driven by HCNM for gene-ranking, integrative subspace identification, and cancer subtype prediction.

1.3 Organization of the Thesis

- Chapter 1 provides the introduction of this thesis, motivation, and proposed workflows that uses correlation networks of biomedical data.

- Chapter 2, the first part, briefly summarizes the brain connectome and different connectivity measures considered in the literature for non-invasive brain modalities such as MRI, electroencephalography (EEG), magnetoencephalography (MEG), positron emission tomography (PET), single-photon emission computerized tomography (SPECT), etc. It presents a brief literature survey of correlation measures used to construct brain functional connectivity networks.

  The second part of this chapter briefs on gene association studies of high throughput sequencing data. It also presents the relevant literature of genomic studies considering multi-omics associations, integrated multi-omics studies, and multilayer networks generated using omics data.

- In Chapter 3, we introduce the resting-state brain FCN. We describe the details such as choice of data, construction of the network, etc. We find modular brain regions utilizing complete/full network using EFA method and compare our results with the state-of-the-art methods.

- Chapter 4 describes the multiscale consensus method to find modularity maximized communities and tight-bound cliques in the resting-state brain FCN.

- In Chapter 5, we introduce multi-omics of cancer data. We propose to use a two-level integrative model: i) a network-free (NF) regression model, and ii) an HCNM. We address the ‘small n large p’ problem of multi-omics by utilizing the multi-level integrative procedure followed by gene-pair ranking, which results in a representative integrative subspace.

- In Chapter 6, we utilize the representative integrative subspace of multi-omics data, and employ a network-based (NB) integrative model to find the subtypes of patients of two cancer profiles a) the most widely studied subtypes, i.e., breast cancer, and b) the least studied subtypes, i.e., lung cancer.

- In Chapter 7, we introduce the extended applications of our studies.
  (*i*). Study of patterns in resting-state FCN, before and after treatment of an obsessive-compulsive disorder (OCD) subject.
  (*ii*). A graph layout 'RadTrix' to visualize an unbalanced bipartite graph.

- Chapter 8 summarizes this thesis with our findings, limitations, and provides the scope of future work.

CHAPTER 2

CORRELATION NETWORKS OF BIOMEDICAL DATA

Biomedical data are measurements gathered from human subjects to study healthy state or illnesses in the body. Statistically analyzing such datasets gives valuable information of patterns observed in a cohort or sub-population with respect to a specific health condition, such as cancer, or in a control group of healthy subjects. Network science provides powerful tools to study statistical relationships between entities in the biomedical datasets, of which correlation is widely studied. In this thesis, we consider correlation studies conducted in brain connectome and multi-omics data in oncology, using relevant biomedical data collected of specific cohorts.

2.1 Brain Connectome

Brain connectivity studies comprise static, dynamic, or causal connections based on their anatomical, functional, or effective connections, respectively. Last two to three decades, non-invasive and in vivo studies are popular and are most widely used due to increased neuroimaging modalities such as EEG, MEG, fMRI, PET, and SPECT. Also, the emerged discipline, 'neuroinformatics,' gave enormous scope for brain connectivity studies. The "connectome" is a human brain that is treated as a connection matrix (adjacency matrix) [48], with rows and columns comprising elements and corresponding interconnections represent the strength of connectivity. The brain connectivity analysis is of three different types, a) structural connectivity (SC), b) functional connectivity (FC), and c) effective connectivity (EC).

- SC is studied using anatomical/physical connections. The connectivity refers to the presence of white matter, fiber tracts that are physically interconnecting different brain regions. These anatomical connections are commonly studied using diffusion magnetic resonance imaging [49–51]. The structural connections are persistent on short time scales, spanning a few seconds and minutes, and brain plasticity changes are observed when examining longer time-span recordings.
- FC is inferred from correlations between nodes (defined for the network [52]) based on the blood-oxygenation level-dependent (BOLD) signals of fMRI imaging or electrical activity of EEG or magnetic activity of MEG and, etc., signals [53, 54] obtained while the subject is at a resting state or performing any cognitive task. The connectivity matrix of functional networks could be computed using several methods [55], e.g., correlation, mutual information, covariance, coherence, etc.

  The focus of this thesis is correlation networks, and our case studies are derived from resting-state (rs-fMRI) data using Pearson's correlation coefficient. While EEG depends on the number of electrodes, the measurements need not cover the entire brain volume. We tested the network of the entire volume, which covers both MEG and fMRI, where we have used fMRI, as FCNs are widely studied.
- EC is the study of causal dynamics that consider the effect of one neuronal system influence on another neuronal system [27,56] and is measured using MRI, EEG, or MEG signals. EC gives the directed causal relationships, e.g., Granger causality among distributed responses. As the 'cause and effect' details are studied here, the connectivity matrix gives a directed network.

These three connectivity matrices can be analyzed as networks. Both SC and FC form an undirected network, whereas EC is a directed network due to the causal relationship between the elements. The SC network is relatively stable compared to FC and EC. Both FC and EC are mostly studied for spontaneous dynamic connections at rest/no-task or performing any cognitive activities. Both EC and FC supplement the knowledge of human brain anatomical connections [54]. Though FC and EC networks are derived using different definitions, the same imaging modalities can be used for both studies [27]. In this thesis, we study brain *FCNs* derived from *rs-fMRI* data and the connectivities are computed using *Pearson's correlation* measure.

2.1.1 Functional Connectivity Networks

The elements that can be treated as nodes in the FCN can be viewed at different brain resolution levels. At a 'macroscopic scale', the anatomically distinct regions, i.e., parcels of remote brain regions are considered as nodes. At a 'mesoscopic scale', the neuronal population level is treated as nodes, and at a 'microscopic scale', each neuron is considered a node. For constructing the connectivity network, the definition of node plays a critical role. Nodes can be of voxels [57], considering each voxel of imaging data as a node or anatomical brain parcellations [58,59] or cytoarchitectonic information based [60]. Stanley et al. [52], in their work, have concluded that, to uncover the novel brain properties, nodes of voxel-based are preferred over anatomical atlas when employing functional activation meta-analytic approaches. The choice of nodes is highly task-specific, and anatomical brain areas can be examined at various resolutions owing to the brain hierarchical modular organization [61]. Korhonen et al. [62], have concluded that the anatomical brain areas that are presumed to be functionally similar are not consistent and the choice of nodes must be carefully decided, as nodes consistency scores varied widely in their experiments. The links or strength of connectivity between these nodes of any scale are computed based on the problem at hand and the relationships between the nodes. In this thesis, we study the resting-state brain connectivity network derived from rs-fMRI data at macroscale and functional connectivities are measured using Pearson's correlation.

Brain functional connectivity network is highly studied under three broad categories [63, 64]: a) Functional segregation, which refers to the extent a network can be segregated/divided to form triangles or cliques or communities. b) Functional integration, which is a global measure and is used to study the flow and exchange of information among the communities by measuring the path lengths of the connectivity networks. c) Functional influence, refers to an individual node or edge contribution to the flow of information in the network and is studied by measuring the centrality scores and finding hub nodes.

Our work focuses on *macro-scale functional connectivities*, to study the non-overlapping communities, i.e., *functional segregation* of the network derived from rs-fMRI (also known as task-free fMRI (TF-fMRI)) data. Functional segregation implies "neuronal processing" carried out in modules containing functionally related brain regions. The regions of interest (ROIs), i.e., nodes of the FCN, exhibit higher interconnection edge density among functionally related regions. This trend allows the FCN to have a significant clustering coefficient (CC). A higher value of CC of a network means, the node's neighbours are neighbours to each other [65], and FCN has exhibited small-world characteristics [66,67]. Rubinov et al. [63] have discussed about *modularity* in networks as a measure of functional segregation. Modularity is a measure to learn the extent a network can partition into non-overlapping communities. While several studies have performed clustering directly on the preprocessed fMRI image data, such as in [68], many have been performed on the network constructed from correlation matrices, e.g. [69].

A large body of work has been focused on analyzing the FCN as a graph or as a network [35]. The nodes of these networks are regions in the brain from a specific parcellation method, e.g., Automated Anatomical Labeling (AAL) [58], Dosenbach atlas (DOS) [70]. Arslan et al. [71] have compared various parcellations that are of, a) anatomical based, b) connectivity-driven, and c) random parcellations. In their work, it has been observed that the task-free or resting-state fMRI data displays better agreement with connectivity-driven parcellations compared to random parcellations and anatomical parcellation methods. The edges between the nodes for functional connectivity are computed based on relationships between different brain regions, which encode the connectivity between the nodes, e.g., correlation, mutual information, phase coherence, partial correlation, spectral coherence, etc. [35, 72]. For example, the network datasets provided in [2, 69] are computed as correlation matrices. Functional connectivity is inferred from correlations between nodes based on the blood-oxygenation level dependent (BOLD) signals in fMRI imaging [53, 54].

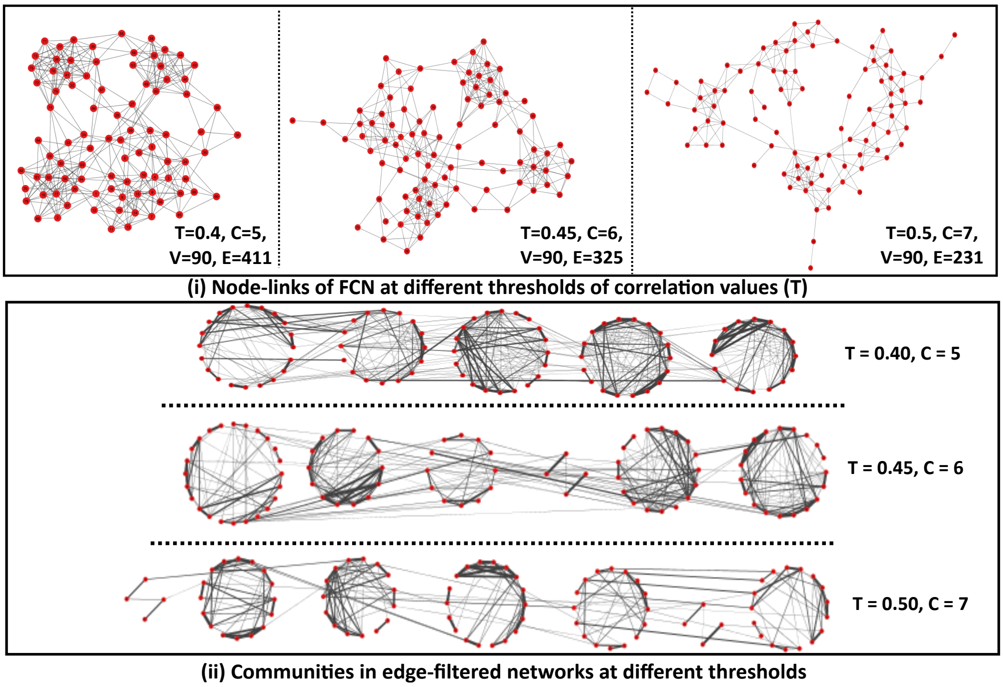

T=0.4, C=5,
V=90, E=411
T=0.45, C=6,
V=90, E=325
T=0.5, C=7,
V=90, E=231
(i) Node-links of FCN at different thresholds of correlation values (T)
T = 0.40, C = 5
T = 0.45, C = 6
T = 0.50, C = 7
(ii) Communities in edge-filtered networks at different thresholds

Figure FC2.1: Network visualization of AAL-90 nodes functional connectivity [2] of fMRI scans from Beijing Normal University in the 1000 Functional Connectome Project [3], at different thresholds of correlation values, T. The number of vertices V indicates the entire network is considered here, while the number of edges, E, decreases and the number of communities, C (extracted using Louvain community detection), increases with the T value. (i) Shows the node-link diagram, (ii) shows the same network with stacked circles with the circular layout of nodes in each community. The edge widths are proportional to the correlation value.

In the conventional workflow of functional network analysis [35–37], the connectivity matrices[1] are subjected to edge-reduction using graph filtering, i.e., filtering out edges below a chosen threshold of edge weight, or outside of a chosen interval(s) of edge weights in the network. The edge-filtered weighted matrices are either used directly as weighted graphs or binarized to give unweighted graphs. Subsequent to graph filtering, the workflow usually progresses with network science-centric and/or graph-theoretic approaches. While thresholding facilitates the reduction of edge density of the network by retaining only significant edges, there also exists the downside that the choice of threshold influences the network topology and the subsequent analysis of strongly connected components [35]. We can observe this in figure FC2.1 as slight changes in thresholds can give different network topologies[2]. The qualitative analysis using Figure FC2.1 with the force-directed graph layout and circular layout of the communities depicts dense inter-community links at threshold 0.4. But, as we increase the threshold, both intra- and inter-community links begin to become sparse, and the FCN segregates to multiple subgraphs leading to over-fragmentation. Hence, it is hard to decide on finding an optimal threshold. Moreover, there is no consensus on the science behind the choice of thresholds [36]. Hence there is a need to study the brain FCN without sparsification.

Studying functional segregation of connectivity networks has had strong interest from the research community, as this allows to learn the spontaneous dynamic patterns in the human brain. Here, network segregation can be inferred from communities/clusters/modules/subgraphs of nodes that exhibit high intra-edge density. Power et al. [73] used rs-fMRI data, with correlations as functional connectivity measure and Infomap [74] algorithm for finding sub-graphs. In this work, the functional segregation of the network is studied as sub-graphs that are in greater agreement with the known

[1] Connectivity matrix of a network is also the adjacency matrix of the graph data structure storing the network data. Similarly, the nodes and links in the network correspond to the vertices and edges of its graph, respectively.

[2] The FCN used here is cross-correlations (Pearson's) among AAL anatomical atlas of 90 nodes, where data is acquired from rs-fMRI modality from healthy right-handed subjects of 18 to 26 years age group.

functional human brain networks. Meunier et al. [61] studied hierarchical modularity in the human brain, where rs-fMRI data is used, and functional connectivity is inferred by wavelet correlations. In this work, the FCN that is thresholded and binarized is subjected to graph-theoretical analysis to verify the hierarchical modularity in the human brain FCNs. Partial correlations between the ROIs of task-based fMRI data are used in [75], where simple finger movement activity/task of the subjects are studied using Bayesian analysis and observed the hemispheric symmetry among the nodes of cortical motor regions. The MRI scans of each subject of this study are the club of both 'rest', and 'activation' (task) states alternatively. Apart from correlation measures for generating FCN, few other measures such as coherence [76] or multivariate mutual information (MI) [77] or multivariate Granger causality (G-causality) [78] are used. Mutual information or coherence is majorly employed where negative values are of concern for brain functional connectivity analysis. In relation to FCN, "Resting State Networks" (RSNs) can be considered as a sub-network of the FCN. De Luca et al. [57] have used probabilistic independent component analysis (PICA), which finds the independent components in the fMRI images, where the components being statistically independent spatial maps, correspond to different activation patterns. Sporns [79] surveyed various community detection procedures focused on finding modules of brain connectivity networks. The methods include modularity maximization, distance-based procedures, simulated annealing, divisive algorithms, spectral decomposition, stochastic block models, greedy algorithms, Infomap, overlapping communities, and independent component analysis (ICA). Though there are various functional connectivity measures, studies often use Pearson cross-correlations for functional connectivity measures [79, 80].

The spontaneous dynamic changes in functional connectivities in healthy controls are regularly supported to study different neurobiological diseases. Jones et al. [28], in their work, have used task-free fMRI data of a large population (892 subjects), who are of advanced age group (70-90+ years) but are cognitively normal (CN). The FCN is generated with 68 nodes/ROIs derived using ICA, and the links of the network are Pearson's correlation coefficients. Using full network (unthresholded) and network-theory based measures of brain connectivity toolbox (BCT) [63], the modular brain organization is first studied in the CN group, and the patterns in Alzheimer's disease (AD) cohort of 28 subjects are verified. Significant differences in anterior and posterior default mode networks, i.e., *aDMN* and *pDMN* are observed. Similar studies are carried out between case-control networks and by examining the spontaneous dynamic changes of functional connectivity in the brain. Significant patterns and differences are noted in neurological diseases such as schizophrenia [29, 30], obsessive compulsory disorder (OCD) [31], attention-deficit/hyperactivity disorder(ADHD) [28], Alzheimer's [28, 32, 33], parkinsonian syndrome [34], etc.

Overall, in this thesis, we study brain functional connectivity networks derived from rs-fMRI data. In our work, the macro-scale functional connectivities are measured using Pearson's correlation. The complete/full FCN is used to examine the functional segregation by finding non-overlapping communities and cliques within communities of the functional connectivity network.

2.2 Multi-omic Associations for Cancer Data Analysis

Cancer is due to the uncontrolled and abnormal proliferation of cells in the organism's body. The recent development of high throughput sequencing (HTS) technologies and next-generation sequencing (NGS) platforms, offering researchers an enormous opportunity to study cancer data at genomic, epigenomic, transcriptomic, and proteomic levels, and as well as integrated genomic analysis. Finding cancer-related genomic abnormalities is much needed for the early prognosis of the disease and treatment.

2.2.1 Correlation Analysis of Cancer Genomic Data

The cancer genomic[3] data is made available through comprehensive sequencing of genomes of initiatives such as The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC) [82], and Pan-Cancer Analysis of Whole Genomes (PCAWG) [83][4], etc. TCGA is supported by the National Cancer Institute and the National Human Genome Research Institute, which had brought together researchers from various fields of many institutions. TCGA data portal comprises a set of 20,000 samples from 33 different cancer types. A large amount of data (over 2.5 petabytes) that is of different genome levels such as DNA methylation, gene expression, protein expression, DNA copy number variation, miRNA expression, and somatic mutation are available publicly at the TCGA data portal. ICGC Data Portal is with 22,330 donors molecular data from 86 different cancer projects (including few TCGA projects) involving 22 cancer primary sites. Most of the data collected at TCGA and ICGC projects is of 'whole exome sequencing' (WES), whereas PCAWG is focused on 'whole genome sequencing' (WGS) data obtained from TCGA and ICGC of various cancer profiles. In this thesis, we focus on studying cancer data that is retrieved from the *TCGA* data portal.

In cancer data analysis, different correlation measures are used to study the effects of omic features, for example, to study the role of DNA methylations that regulate the expression value of oncogenes. In general, DNA methylation plays a significant role in blocking transcription factors from binding, leading to a decrease in gene expression value [84]. The associations between DNA methylations and gene expression data of prostate adenocarcinoma are studied using correlation analysis and found novel cancer biomarkers and hub oncogenes that are critical for prostate cancer [85]. Similarly, Spainhour et al. [86] have used TCGA data for 33 different cancer types, and

[3] The word 'genomic' of an organism refers to, study of genes, their influence, interrelations, and functions [81].

[4] The data portals used to query, download and visualize the cancer data of TCGA, ICGC and PCAWG are i) `https://portal.gdc.cancer.gov/`, ii) `https://dcc.icgc.org/`, iii) `https://docs.icgc.org/pcawg/` respectively.

have argued that, methylation and gene expression correlation (Pearson's) patterns are mostly tissue independent, and the role of DNA methylation is not confined to the silencing of gene expression. In few other works, methylation and gene expression correlations are studied to identify crucial genes in severe oligozoospermia disease [87], breast cancer [88], bipolar disorder, and schizophrenia [89], etc. Apart from Pearson's correlations, studies also used measures such as weighted correlation network analysis (WGCNA) [90–92], partial correlation [93–95], multivariate multiple (linear) regression (MMR) [24, 41], etc., to investigate the gene-gene associations of methylation and gene expressions of various disease datasets. The correlation/association studies are not limited to methylations and mRNA expressions (messenger RNA). Studies also consider associations between miRNAs (microRNA) and mRNAs to predict the mRNA genes targetted by miRNAs, for a specific biological context. A known phenomenon to observe here is the down-regulation effect of miRNAs on its targets, i.e., if the expression of particular miRNA increases, then that of its target mRNAs decreases. Aziz et al. [96] have used regularized regression to construct miRNA and gene expressions network. Similarly, Muniategui et al. [97] have proposed Target Lasso (TaLasso) regression analysis, Iterson et al. [98] and Lee et al. [99] have used integrated approaches to study miRNA–mRNA associations. There also has been work in using cross-correlations between different omics features for gene-gene interaction networks.

For cancer studies, integrating multi-omics data enables in recognizing biologically significant interpretations that facilitate enriched cancer outcome predictions [42, 44, 100]. Integrative multi-omics based studies are well-known in recent years [43,101]. Integrative association studies between disease-genotypes, e.g., DNA methylation, miRNAs, copy number variations (CNV) and mRNA expression levels, with respect to disease-phenotypes, e.g., different cancer profiles, are important for making sense of genomic datasets. Kim et al. [102] have addressed the need for integrating multi-omics data, which leads to specific cancer outcome prediction. Ding et al. [103] have used visual analytics of stratification of patient data in TCGA to integrate multi-omics data.

The gene-gene associations either of single- or multi-omic are treated as networks and studied further using network-theory measures. Networks are crucial for understanding various phenomena in biology, as these networks have the distinction of the feasibility of validation using alternative yet standard methods in the domain. Barabási et al. [104] have stressed on the systems-based network analysis for understanding interconnections between human diseases. Today, genome-wide studies provide rich data for identifying these interdependencies, which are captured best, using networks. Gosak et al. [105] have reviewed the use of multilayer networks (MLN) as an effective model of complex interactions in biological datasets, e.g., the interdependencies between diseases using separate phenotype and genotype layers. De Domenico [106] has further reinforced the need for MLN in multi-omics data to discover new insights into the evolution of the disease. However, the lack of a "consistent, replicable, and reproducible" MLN model remains an open challenge in multi-omics [105, 106].

While there has been work in specific integrative analysis, there has been a limited effort in combining some of the existing analyses to build effective data science workflows. In this thesis, we use the multi-omics data of cancer profiles downloaded from the TCGA data portal. We study the genes and their associations using a heterogeneous correlation network model, which is a type of multilayer network. We use this model to bring together different existing integrative approaches.

Part I

Brain Functional Connectivity Correlation Networks

CHAPTER 3

MODULARITY MAXIMIZED COMMUNITIES IN BRAIN USING FACTOR ANALYSIS

The communities/modules in brain FCN are sub-networks of the brain with dense intra-community connections and sparse inter-community connections. Nodes of the network in each module help understand the functional behavior and dynamic connections of the resting-state functional brain network. Multiple studies have shown the presence of modules in the brains of different organisms. For example, the Caenorhabditis elegans (*C. elegans*) which is the only species with 302 identified neurons and a respective neuron map [107]. Several network-centric methods, including the modified modularity maximization approach for hierarchical community detection [108], and the stochastic block modeling [109], have been effectively used to extract the community structure in the *C. elegans'* brain connectome.

The functional segregation of the brain network is synonymous to a community in the network-science that refers to neuronal processing carried out among communities. Segregation refers to cliques or communities or motifs or the number of triangles of a network, which can explain the extent a network can form into separate components [24]. We use a brain FCN that contains pairwise correlations of the nodes, where the nodes are anatomically derived parcels of the brain at a macroscopic scale. Here, we express how we can exploit the semantics of the adjacency matrix and use matrix-based

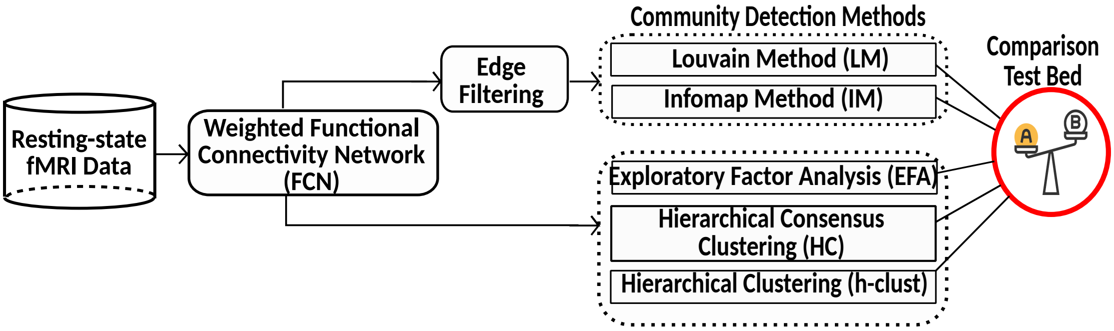

Community Detection Methods
Resting-state fMRI Data
Weighted Functional Connectivity Network (FCN)
Edge Filtering
Louvain Method (LM)
Infomap Method (IM)
Exploratory Factor Analysis (EFA)
Hierarchical Consensus Clustering (HC)
Hierarchical Clustering (h-clust)
Comparison Test Bed
A
B

Figure FC3.1: Workflow for comparing various community-detection/node-partitioning techniques that use a combination of methods that utilizes a full and sparsified brain functional connectivity network.

approaches for brain functional network analysis. Our approach is different from widely used network-based procedures that require network sparsification, which is achieved by applying a threshold value on the edge weights. In the traditional workflow of functional connectivity network analysis [36, 37], the connectivity matrices are subjected to sparsification, i.e., filtering edges using a chosen value of the edge-weight threshold. These edge-filtered, weighted matrices are either used directly as weighted networks or binarized to provide unweighted networks. The thresholding process has been debated in the literature [35], as the choice of threshold value influences the network topology (Figure FC2.1). The filtering process discards weaker edges irrespective of their relevance for functional connectome. Hence, in this study, we use a complete network which is an FCN and EFA procedure, to functional segregate the network.

In this chapter, we find the modularly organized brain communities[1] in the FCN using EFA, compare our results with the state-of-the-art methods, and also compare with the results of relevant literature [110]. We propose a case study to compare these methods, as depicted in Figure FC3.1.

The following sections of this chapter describe: i) our proposed method for functional segregation (Section 3.1), ii) FCN generation (Section 3.2), iii) edge filtering pro-

[1] We use the terms communities, node-groupings, node-partitions, clusters, modules interchangeably.

cedure (Section 3.3), iv) community-detection/Node-partition methods (Section 3.4), and v) comparison of the node partitions (Section 3.5).

3.1 Proposed Approach for Functional Segregation

For computing brain functional segregation of the resting-state, weighted, fully connected, and undirected FCN, we propose to use EFA. Charles Spearman [38] first proposed factor analysis; in the domain of psychology study. Spearman demonstrated that there exists a reason or common factor called ‘general intelligence’ for the grades in school subjects such as maths, history, etc. EFA is predominantly used to study the structure of the variables. EFA can be of R-type or Q-type analysis of a correlation matrix [111, 112]. The R-type factor analysis finds the latent structure of the variables, whereas the Q-type factor analysis uses a subject-wise correlation matrix and is used to cluster subjects in a population. While Q-type analysis has been used for community detection [113], R-type factor analysis is apt for functional segregation of the FCN with maximized modularity. EFA is an exploratory and experimental method used to find the underlying structure of the data [112]. Section 3.4.1 elaborates the details of EFA.

3.2 FCN generation

Datasets for Case Study: The resting-state functional connectome dataset, i.e., fMRI scans of the healthy right-handed volunteers of young adults of age group 18-26 years, is used to generate FCN. MRI is the most popularly used modality for human brain studies. Modeling the MRI data as functional connectivity networks culminates in network-scientific analysis [35, 64, 114]. The dataset is from Beijing Normal University, from the 1000 Functional Connectome Project [3]. The dataset is of 198 subjects of right gender-balanced (122 female volunteers). Data acquisition is made with the Siemens

3T scanner, and data preprocessing is done using SPM5 (Statistical Parametric Mapping) and DPARSF [2]. The fMRI scans of each subject are captured while the subject is in resting-state with an eyes-closed (EC) condition yet being awake. The Automated Anatomical Labeling (AAL) parcellation atlas [58] of 90 ROIs of the brain cerebrum is used to define the nodes of the network. Note that the ROIs of the cerebellum of the brain are excluded in this FCN generation. In this work, the network is addressed as $\mathbf{AAL}_{90}$, referring to the parcellation method and its number of nodes.

Functional Connectome Generation: For each subject data in the cohort, an FCN using fMRI data is generated from the extracted mean time courses (BOLD signal) of the nodes/ROIs. Pairwise, Pearson's correlation among all the nodes gives an undirected, weighted, and completely connected FCN. Fisher's r-to-z transformation is applied on each correlation network and derived z-score matrices. The final 'correlation matrix' is generated by aggregating individual FCN matrices, which is an adjacency matrix. $\mathbf{AAL}_{90}$ is an adjacency matrix of size 90, that uses a 90-node AAL parcellation. The final $\mathbf{AAL}_{90}$ FCN is subjected to different community detection algorithms for the comparative analysis of functional segregation.

The methods Louvain (LM) and Infomap (IM) of our comparison test-bed are not applied on full FCN; hence we find an optimal edge-weight threshold value to sparsify the network. The rest of the methods, i.e., Exploratory Factor Analysis (EFA), Hierarchical consensus clustering (HC), and Hierarchical clustering (h-clust) are operated on full FCN, but these methods expect an input parameter that specify the number of clusters (k) to be derived from the given network. Hence we need to find an *optimal threshold* value to filter the network and the *optimal number of communities* that can produce the communities with the maximized modularity.

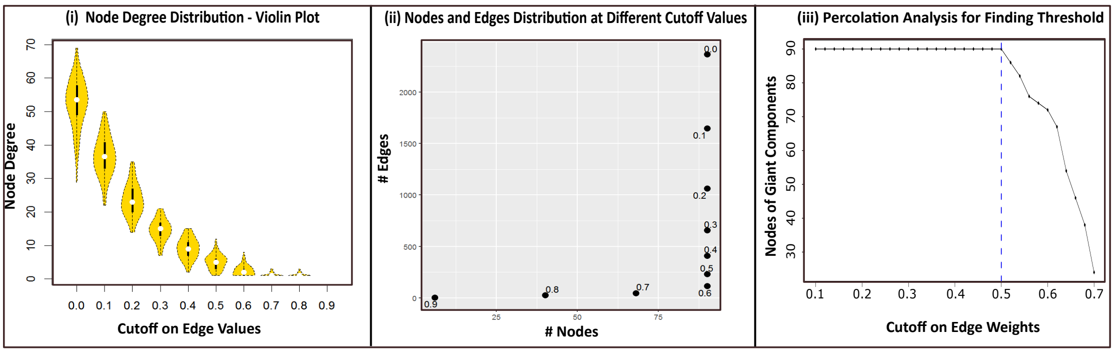


Figure FC3.2: Finding edge threshold for $\mathbf{AAL}_{90}$ network. The figure depicts: (i) violin plot of node degree distribution at different thresholds of edge-weight and the elbow graph to find the optimal threshold (T) value, (ii) the number of nodes vs. number of edges plot for different thresholds, and (iii) the number of nodes of giant components at each edge threshold using percolation analysis.

3.3 Edge-Filtering Networks

Most network-science based FC studies had considered edge-filtering as a preprocessing procedure. The cut-off value for edge reduction can be realized by observing the network properties such as nodes degree distribution, edge distribution, etc., at each increasing value of edge-weights.

To sparsify the FCN and to implement Louvain and Infomap community detection methods, the threshold value is identified by studying: (a) the network node degree distribution using an elbow curve (Figure FC3.2 (i)), (b) by observing the number of nodes vs. the number of edges at each edge-weight threshold (Figure FC3.2 (ii)), and (c) by verifying the number of nodes of largest connected components as we remove edges in decreasing order of edge-weights using percolation analysis [115] (Figure FC3.2 (iii)). In percolation analysis, an edge-weight value is considered a final cut-off value, when at values higher than that edge-weight, the network tends to break down into a larger number of smaller subnetworks. Figure FC3.2 (ii) shows a sharp drop in the edges until a threshold of 0.45, beyond which the number of nodes reduces. Similarly, in Figure FC3.2 (iii), the number of nodes of the giant connected component was constant until a threshold value of 0.5; beyond this value, the nodes start decreasing as the network starts fragmenting into multiple connected components. The node degree distribution in Figure FC3.2 (i) indicates a range of values, i.e., 0.4, 0.45, and 0.5. Hence, in this study, for **AAL**$_{90}$ FCN, we consider edge-weight threshold values as $\tau = \{0.4, 0.45, 0.5\}$.

3.4 Community-detection/Node-partition Methods

This section first elaborates our proposed functional segregation procedure, i.e., EFA, followed by a few carefully chosen state-of-the-art community detection algorithms that are used to compare EFA results.

3.4.1 Exploratory Factor Analysis (EFA) for Node-Partitioning

We *partition* the FCN by identifying factors corresponding to node-partitions in the network. The factor analysis works with an assumption that there exists an underlying structure among the variables. If the correlation values between the variables are not significant, then EFA fails to identify groups of variables, owing to the absence of structure among the variables. If these groups are located, they exhibit homogeneous distribution [111]. The nodes of the FCN are treated as random variables, thus, allowing the use of EFA to group nodes. *Functional segregation* refers to the modules of a network with nodes that are functionally related and tightly connected due to homogeneous edge distribution. Here, we use EFA to determine the functional segregation of the FCNs, *i.e.*, to locate the strongly inter-linked nodes [41].

Factors in EFA can be computed using either *component factor analysis* or *common factor analysis* method, based on the type of variance considered among the variables. The variance of any variable can be divided into three parts, i.e., common variance, specific or unique variance, and error variance [111]. Component factor analysis methods use the total variance of the variable, *e.g.*, the principal component method. In comparison, methods such as principal axis factoring and maximum likelihood use common factor analysis, where only common variance among the variables are considered to identify an underlying structure that leads to differentiate factors. The unique and error variances are not associated with correlation values among the variables but are due to unreliability while collecting data or measurement error.

Here, we use the *maximum likelihood estimation (MLE)* method to identify factors. MLE is highly recommended amongst all FA extraction methods [116], as it provides relatively more information about the factors, and it is the best-suited method when data is normally distributed. A $[p \times m]$ matrix is the *factor loading* matrix is computed using MLE for p observable variables and m reduced latent, unobservable variables that share similar variance.

Factor loadings are the correlations of variables with the factors. To attribute labels and for the right interpretation of the factors, orthogonal or oblique rotation methods are used in factor analysis. The rotation procedure can decrease the ambiguity in the solution and aid in labeling the factors. Multiplication of the loadings with an orthogonal matrix, which is equivalent to a rotation (i.e., a linear transformation), does not change the covariance matrix that is regenerated from the transformed loadings [117]. This transformation does not modify the communalities. *Communality* is the measure used in EFA to determine how well the node correlates with others, based on the number of factors used. Communality greater than 0.4 is desirable for all variables for any number of factors in EFA [118, 119]. One way to find a consistent set of factors using factor loadings (rows of $[p \times m]$ matrix) is to find an appropriate rotation that can allow the interpretation of factors and also maximize the communalities using a "simpler structure" [120]. Popularly used rotation methods are *varimax* rotation, an orthogonal, and an *oblimin* rotation, non-orthogonal.

Our four-step algorithm for using EFA on the FCN involves: (1) checking the feasibility for implementing EFA on the corresponding correlation matrix, (2) estimating n_F, (3) implementing EFA using MLE with both rotation methods, and (4) retaining node-grouping from communality-maximizing rotation method. Elaborated details of EFA are provided in *Appendix A*.

Eligibility of Correlation Matrix for EFA: For finding factors of a correlation matrix, the primary eligibility criterion is that the correlation matrix has to be positive definite to ensure non-singularity. However, it must be noted that a few of the EFA methods and implementations relax the criterion to the property of positive semi-definiteness. The Kaiser-Meyer-Olkin's (KMO) test is an important test on the correlation matrix to check if its measure of sampling adequacy (MSA) is not less than 0.7 [121]. MSA is a measure to verify the variables linear dependency, as FA anticipates the variables dependency. The correlation matrix with MSA less than 0.5 is not eligible for EFA [121,122]. In our work, the $\mathbf{AAL}_{90}$ correlation matrix is positive semidefinite and has an MSA of 0.78. Thus, the $\mathbf{AAL}_{90}$ FCN is eligible to node-partition using EFA.

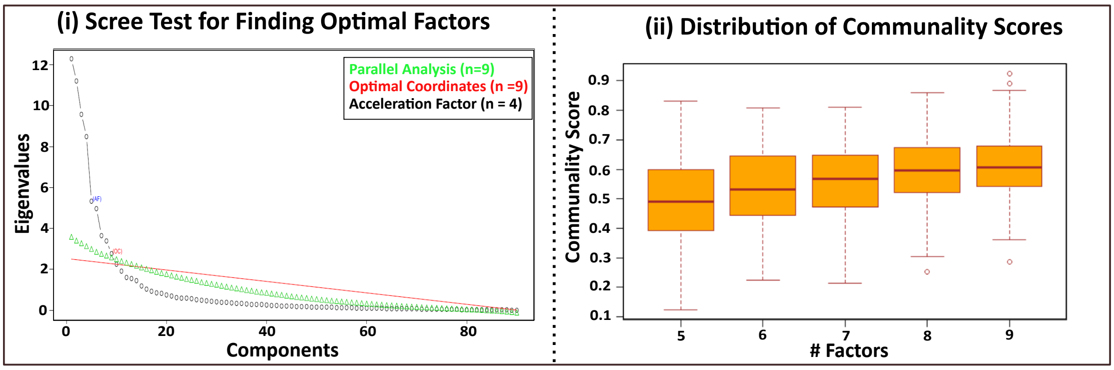


Figure FC3.3: (i) Scree plot to represent the number of factors n_F identified using parallel analysis. (ii) Distribution of communality scores for EFA using maximum likelihood and varimax rotation method at each factor n_F.

Estimation of Number of Factors (n_F): We use variance- or eigenvalue-based methods to estimate the number of factors (n_F) for EFA. A few of the widely used procedures for determining n_F are the *Scree test* and *Parallel Analysis* [123]. The choice of n_F can also be made using prior knowledge of the existing studies on a similar set of random variables [111].

For our case-study data, i.e., $\mathbf{AAL}_{90}$, the optimal number of factors according to the parallel analysis scree plot (n_F^{st}) is *nine*, (Figure FC3.3 (i)). But, there is no rule of thumb to approve or reject any of the values for n_F in EFA. Hence, we have verified a) the number of communities the community-detection methods Louvain (n_F^{lc}) and Infomap (n_F^{im}) produced at the optimal edge-thresholded network (Section 3.3), and b) Number of modules (n_F^{pr}) found in a prior study that had considered a similar dataset and parcellation method, which is investigated by He *et al.* [4].

The number of communities derived on sparsified FCN by Louvain and Infomap methods at 0.5 edge-thresholded network is n_F^{lc} = 7 (Figure FC2.1), n_F^{im} = 12 respectively. But, we have observed that IM had produced over-fragmented sub-networks, i.e., n_F^{im} = 12 at $\tau = 0.5$. Hence we restrict to apply a threshold of $\tau = 0.45$ for method IM, which produces n_F^{im} = 9. The number of modules derived in prior studies is n_F^{pr} = 5, and using parallel analysis, scree test is n_F^{st} = 9. Using the extrema from these values, the determined interval for implementing EFA is **[5, 9]**. Examining the factors for the range of the interval adds to the exploratory and experimental characteristics of EFA. Hence we find factors of the full FCN for various node-partitions/factors starting from 5 factors to 9 factors.

Factor Loadings for Node-Partitions: When a factor is to be correlated with a variable, the strength of this relationship is captured by its factor loading f_s. Conventionally, retaining this relationship is based on satisfying the condition $f_s > 0.3$ [118]. When the condition is not satisfied, the variable does not correlate with any of the computed factors. When the variable is not correlated with any factor, it forms a singleton partition, resulting in an overly fragmented FCN, which is not desirable. Hence, to ensure the inclusivity of all nodes in the FCN in the node-partitions, we relax this criterion to

$f_s > 0$, as implemented in our previous work [41]. Also, as a factor retention criterion (FRC) [40], we use only the factors with at least one node correlated with the factor.

EFA Implementation Parameters:

Table TC3.1: Communality scores of *community-3* nodes of $\mathbf{AAL}_{90}$ network for n_F = 5. Note that for ease of representation, we have provided the scores of a single community.

Node Name	Communalities Scores	
	Oblimin	**Varimax**
CAL.L	0.7014	0.702
CAL.R	0.7144	0.7305
CUN.L	0.4511	0.4654
CUN.R	0.4654	0.4744
LING.L	0.7752	0.7904
LING.R	0.7871	0.7841
SOG.L	0.5548	0.5897
SOG.R	0.5719	0.5919
MOG.L	0.5982	0.627
MOG.R	0.4903	0.4839
IOG.L	0.4462	0.4704
IOG.R	0.3579	0.3649
FFG.L	0.6359	0.6015
FFG.R	0.6384	0.6275

Between oblimin and varimax rotation methods, the varimax rotation method is chosen, as it shows better communalities scores (h^2) than the oblimin rotation (Table TC3.1). A good communality score for the variables is recommended as it describes the common variance of the variables, where unique or specific variance $u^2 = (1 - h^2)$. Conventionally, communality scores in EFA are used to eliminate variables from the study for dimensionality reduction. Variables with communality score in EFA $h^2 < 0.2$ are eliminated for postprocessing [124]. There are several thresholds of communalities score recommended in the literature for retaining variables, *e.g.* $h^2 = 0.4$ [118, 119], and $h^2 = 0.5$ [111]. For our experiments, Figure FC3.3 (ii) shows the distribution of communalities scores at each selected n_F of EFA. We observe that the communality scores are significant, and the median value at each scale is greater than 0.4.

We conclude that *EFA* with the *maximum likelihood method* and with *varimax* rotation produces acceptable error terms and communality scores of the FCN. We study the factors using EFA for the whole interval, $I = [5,\ 9]$.

3.4.2 State-of-the-Art Community Detection Methods

As depicted in Figure FC3.1, *Louvain* (**LM**), *Infomap* (**IM**), *Hierarchical clustering* (**h-clust**), and *Hierarchical consensus clustering* (**HC**) methods are used for finding clusters in the FCN. There are several classes of community detection algorithms [113, 125], of which modularity optimization algorithms will be directly applicable.

The use of modularity as a measure of functional segregation and for finding the modular structure of the brain has been well-studied [63]. Louvain community detection [126] is a widely used algorithm, which has been used in [61] for computing modularity in the thresholded unweighted graph of the functional connectivity matrix. Similarly, LM has been extended for the weighted graph, using a threshold window [127]. In a similar vein, the information-based algorithm, IM [74], is also applicable for thresholded weighted and unweighted networks. Both the methods, LM and IM, are graph-based algorithms and are used for finding communities on the filtered network.

LM utilizes an iterative procedure that considers a greedy optimization process to find the modules with the maximized modularity score (Q) [126]. The method first considers each node of a network as a community, and over multiple iterations, the neighboring nodes are grouped as a community by verifying the Q value. The maximized Q value is achieved when intra-cluster edge density is higher than the inter-cluster edge density, signifying dense intra-community and sparse inter-community edges.

IM is an information-theoretical method, which is the fastest, most frequently used, and accurate method for identifying communities [125] and is often used in FCN. The root principle of IM is; a random walker is likely to dwell and make more hops inside a community than across the communities. Random walker's state of the path can be described using Huffman coding in two phases. The first phase separates the network communities, and the second phase allots the nodes to the encoded communities. With this intuition, IM finds accurate communities in the network. Both LM and IM exploit the network topology and find an appropriate segmentation of the network that results in dense subnetworks. Thus, these methods are semantically different from EFA, h-clust, and HC, which use a full (complete) network for finding communities.

h-clust a data mining clustering procedure, is used to extract hierarchical clusters; owing to the brain hierarchical modular organization [61]. We have implemented h-clust on FCN by considering, single, complete, average, and ward linkage methods of the hierarchical clustering algorithm.

HC is also a hierarchical clustering procedure that is applied on a consensus matrix. HC uses a two-step process to find communities in the brain networks. As a first step, the generalized Louvain community detection (genLouvain) [128] method with fixed or varying resolution parameter (γ) is used for n iterations, and the second step uses the hierarchical clustering method to find clusters on aggregated results of genLouvain. In our work, we have used genLouvain with a fixed resolution, i.e., $\gamma = 1$ for 100 runs and the used HC with $\alpha = 0.1$ (90%) [129], the value of α decides the co-clustering tendency of two nodes of a network, i.e., checking if the two nodes of a cluster are grouped together by a random chance or by their natural clustering tendency. For both h-clust and HC, the tree-cut of the generated dendrogram of FCN decides the number of clusters to examine in the network.

We compare the node-partitions (n_P) obtained using EFA with the communities/-clusters determined using LM, IM, h-clust, and HC procedures [110].

3.5 Comparison of EFA with State-of-the-Art Methods

Using **AAL**$_{90}$ data, we build a test-bed for a comparative analysis of brain node-partitioning. For comparative study, the modularity score (Q) is used for quantitative analysis, and Sankey diagrams [130] are used for qualitative analysis.

Modularity (Q): The modularity is a network measure that is proposed by Newman and Girvan [131] and then extended to weighted networks [132]. The measure Q is derived using *mixing parameter (μ)* measurement [133]. Q measures the node-partitions score using their intra-community edge-density and the expected such edge-density in a random network. The expected edges in the random network preserve the same node degree distribution as the existing network, but the links between the nodes are connected randomly. The fraction of edges of the nodes that belong to the same community is given as,

$$\frac{\sum_{ij} A_{ij} \delta\left(c_i, c_j\right)}{\sum_{ij} A_{ij}} \qquad \text{(Eqn 3.1)}$$

where $\sum_{ij} A_{ij}$ is equal to $2m$, i.e., the total number of edges in the undirected graph, c_i is the community of the node i and δ function is 1 if $c_i = c_j$, otherwise δ function is 0. The fraction of expected edges is, $k_i k_j / 2m$, k_i is the degree of node i. The modularity value Q is:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta\left(c_i, c_j\right) \qquad \text{(Eqn 3.2)}$$

For a network, the value of Q indicates the goodness of the communities division. The value of Q zero or less signifies that the intra-community edges are less than one can expect by a random chance. A positive Q value around 0.3 and above indicates clarity in partitioning [132].

The Q value is observed high in all methods when the network is segmented into five partitions (Figure FC3.4 (i)); this also supports the five modules of prior work n_F^{pr} [4], which had considered a similar dataset and parcellation method. The modularity

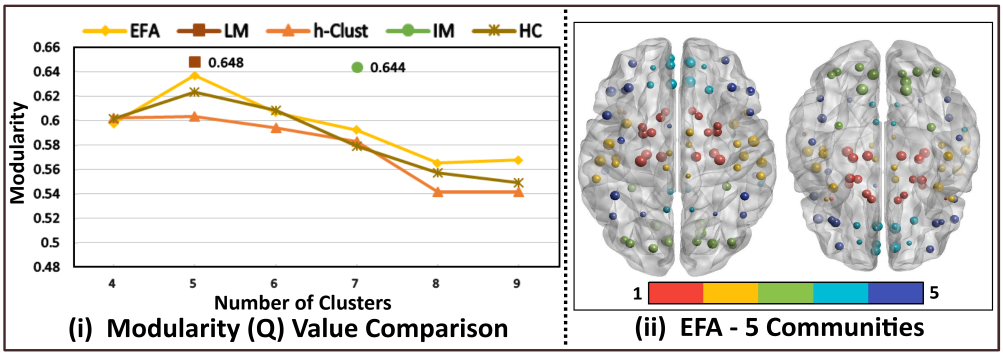
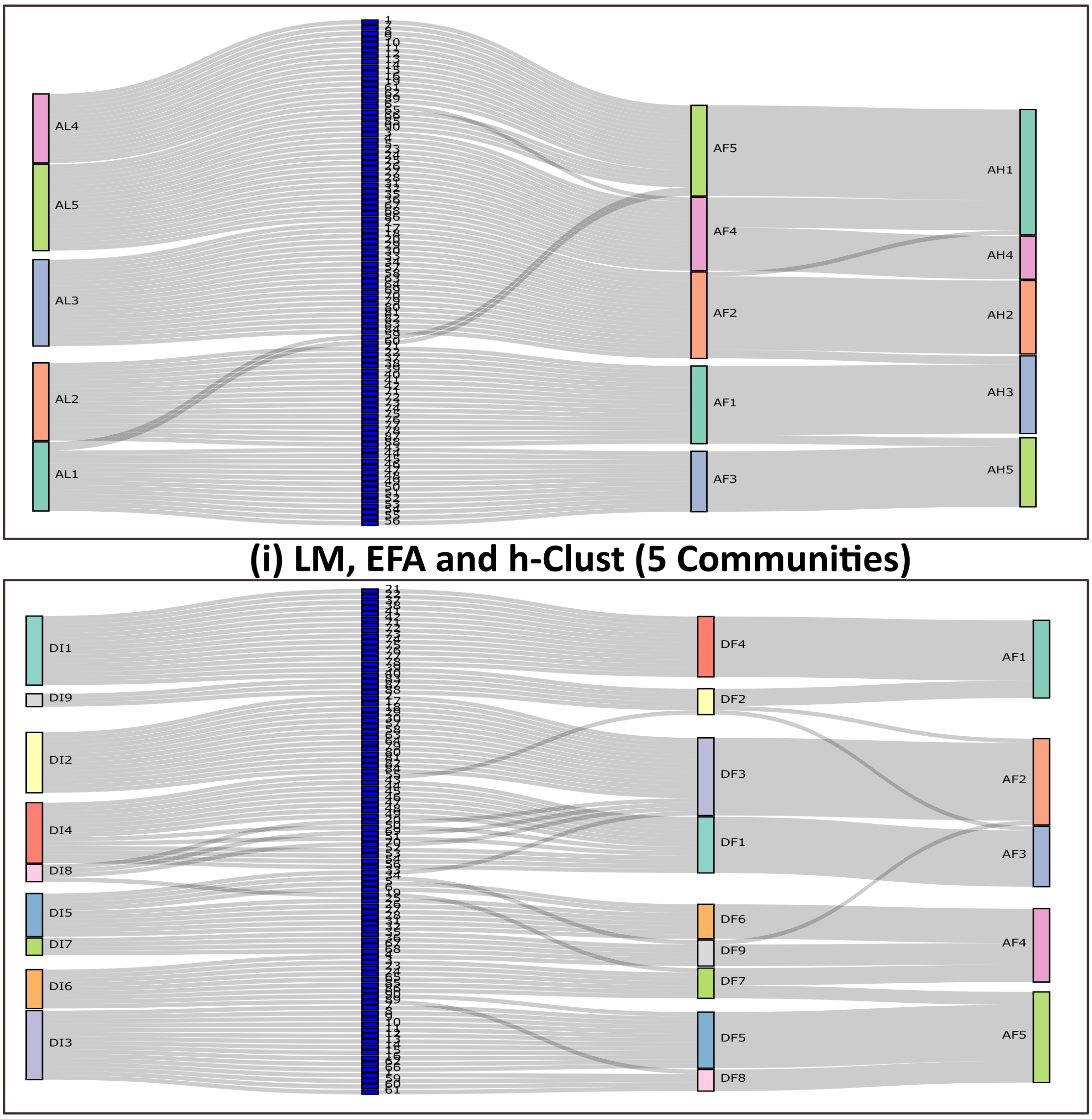


Figure FC3.4: (i) A comparative plot of modularity (Q) scores shows the highest Q value for all methods when the network is partitioned into five communities. (ii) Visual representation of node-grouping of EFA with $n_F = 5$ on the brain surface. The spatial centroid coordinates of nodes are used to display modules in the 'axial view' of the brain, top side view (left), and bottom side view (right). The plot is generated using a MATLAB tool named 'BrainNet Viewer' (BNV) [2].

score of LM is the highest and of EFA is comparable to LM. The highest Q value of the Louvain method can be attributed to its greedy optimization process. The node-groupings of EFA at $n_F = 5$ are visualized on the brain surface, axial view (Figure FC3.4 (ii)). The nodes of each community display modular organization by grouping the nodes in such a manner, where a node's neighbors are neighbors to each other. Along with the modular organization of nodes, hemispheric symmetry is also observed in Figure FC3.4 (ii), i.e., the right and left spatial coordinates of brain regions/nodes are co-clustered in the same community.

The correspondence of communities between the methods LM, IM, EFA, HC, and h-clust is studied using Sankey plots. With the preferred range of edge-weight threshold values of $\tau = \{0.4,\ 0.45,\ 0.5\}$ (Figure FC3.2), methods LM and IM had resulted in $\{5,\ 6,\ 7\}$ and $\{7,\ 9,\ 12\}$ communities respectively. To check the agreement between the nodes when compared with the five modules of 'prior study' (PR) by He *et al.* [4], the five communities of method LM with $\tau = 0.4$ are used. Similarly, the nine communities of the method IM at $\tau = 0.45$ are compared with the node-groupings of EFA identified using the scree test (Figure FC3.3), i.e., with $n_F = 9$.



(i) LM, EFA and h-Clust (5 Communities)

(ii) IM and EFA (9 Communities)

Figure FC3.5: The blue vertical bar refers to node-Ids. The node-groupings are compared between (i) LM at $\tau = 0.4$, EFA with $n_F = 5$ and h-clust using average-linkage method, with $n_P = 5$. (ii) IM at $\tau = 0.45$, EFA with $n_F = 9$ and $n_F = 5$. A detailed explanation of naming conventions used on the Sankey plot is provided in footnote[2].

Figure FC3.5 (i)[2] is the comparison of nodes correspondence of LM with $T = 0.4$ with EFA for $n_F = 5$. We have observed almost similar node-groupings among the communities; 83 out of 90 nodes were grouped similarly in both the methods. Here the edge-crossings, i.e., mismatched nodes correspondence is observed in EFA for a single node at AF4 and six nodes at AF5. A similar proportion of mismatched correspondence is observed between EFA and h-clust methods. For h-clust, the nodes correspondence with EFA is verified for single, complete, ward, and average-linkage methods. The highest matching percentage is observed with the average-linkage method. In Figure FC3.5 (i), the depicted h-clust are resulted by employing the average-linkage method. Though the matching percentage of h-clust with EFA is 86.67%, except for AH2 cluster, the rest all are noticed with edge-crossings that depict differing node-mappings with EFA.

As the method IM at $\tau = 0.45$ and EFA with scree test produces the same number of node-partitions, we have compared their nodes correspondence (Figure FC3.5 (ii)). We have observed high edge crossings in Figure FC3.5 (ii) between the nine communities in IM and nine factors in EFA, indicating failure in the correspondence of node-groupings. Interestingly, the lesser edge crossings are noticed between EFA with 'five factors', i.e., $n_F = 5$ and 'nine factors', i.e., $n_F = 9$. This reveals the hierarchical modular organization of the FCN, which is also a property of the brain connectivity network. The given FCN of $\mathbf{AAL}_{90}$ data, the brain connectivity network exhibited *hierarchical modularity* when n_F in EFA is increased progressively (Figure FC3.7(i)).

The comparison of node-groupings against PR [4] using Sankey plots is shown in Figure FC3.6 (i) to (iv). The findings are: (i) lesser edge-crossings and mismatched node-groupings are noticed with EFA, and the opposite is noticed with the method h-clust. (ii) Nodes correspondence matching percentages with PR, is 90.00%, 88.89%,

[2] For Sankey plots, i.e., in Figures FC3.5, FC3.6, and FC3.7, the *XY* naming convention is used, where *X* is {A, B, C, D}, that corresponds to {5, 6, 7, 9} node-communities. The value of *Y* is {L, I, F, H, HC, He}, which corresponds to {LM, IM, EFA, h-clust, HC, PR}, respectively. For example, **DI7** represents the seventh community out of nine communities (*D = 9*) identified using the method IM (*I*). The automated code to implement the Sankey plots is available at GitHub repository, `https://github.com/vrrani/SankeyPlot`.

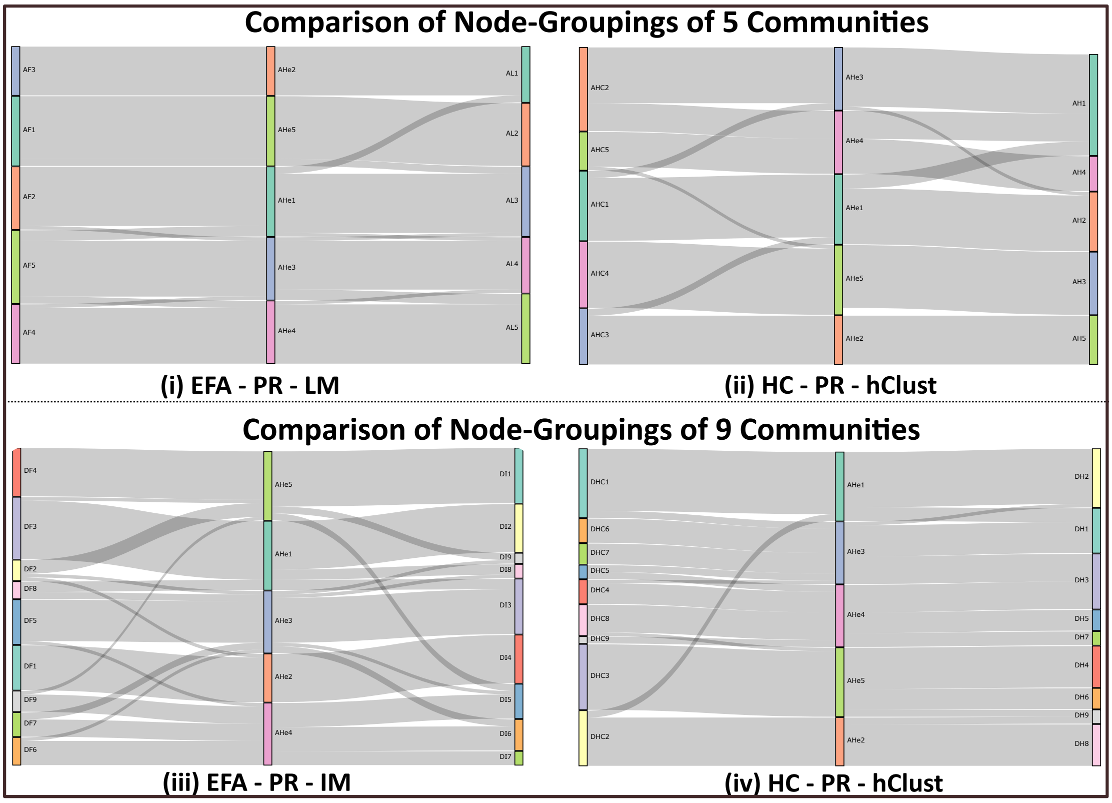
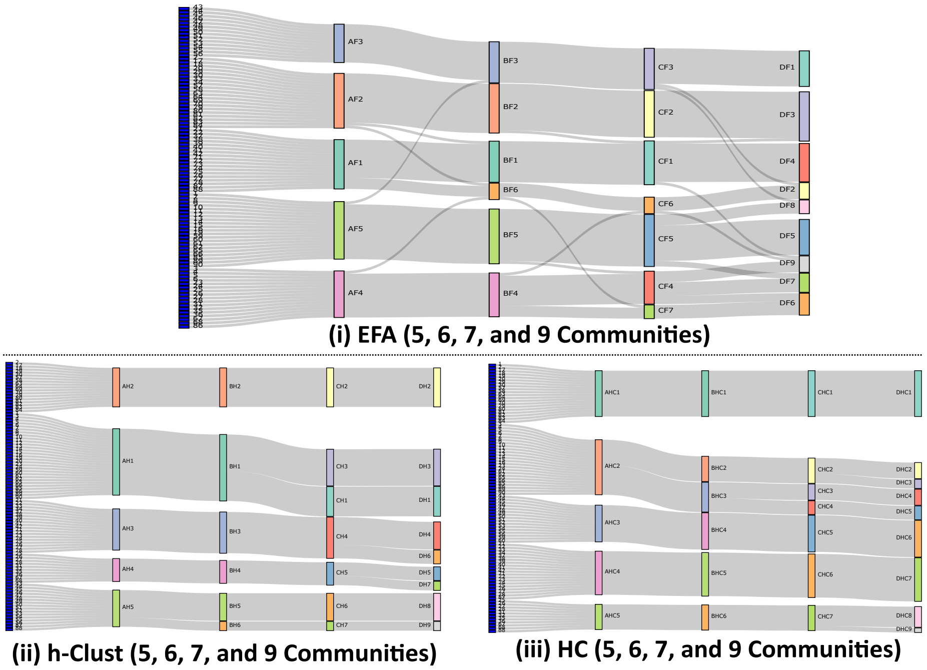


Figure FC3.6: Comparative visualization of nodes correspondence between prior study modules (PR) [4] and node-groupings of our selected approaches. Comparison against PR to (i). EFA with $n_F = 5$ and LM at $T = 0.4$, (ii). h-clust and HC, at $n_P = 5$, (iii). EFA with $n_F = 9$ and IM at $T = 0.45$, (iv). h-clust and HC at $n_P = 9$. The naming convention of the communities is given in footnote[2].



(i) EFA (5, 6, 7, and 9 Communities)

(ii) h-Clust (5, 6, 7, and 9 Communities)

(iii) HC (5, 6, 7, and 9 Communities)

Figure FC3.7: Sankey plot for hierarchical modularity from five to nine modules using the method (i). EFA, (ii). h-clust, and (iii). HC. The naming convention of the communities is given in footnote[2].

85.56%, and 83.34%, for EFA, LM, HC, and h-clust, respectively. (iii) On visual inspection of Sankey plots, the least edge-crossings were observed for EFA-PR-LM (Figure FC3.6 (i)) and HC-PR-h-clust, for $n_P = 9$ (Figure FC3.6 (iv)). On closer observation of the latter plot, we have observed that the size distribution between HC-PR-h-clust is not matched

Within the range of interval $[5,\ 9]$ for n_P, by providing appropriate tree-cut, due to the default nature of hierarchical community detection methods, we observe a clear hierarchy with the h-clust and HC methods Figure FC3.7 (ii). However, interestingly we observe the hierarchical organization of node-groupings with EFA when we run the method for varying number of factors n_F in the range of $[5,\ 9]$ as shown in Figure FC3.7 (i). In the Figure FC3.7 (i), module AF1 is subdivided into module BF1 and BF6 to group the nodes into 6 factors from 5 factors; similarly, module BF4 divides into modules, CF4 and CF7, to group nodes of 6 factors to 7 factors, and a similar pattern of transition is observed from 7 to 9 factors. Though h-clust, and HC (Figure FC3.7 (ii) and (iii)) depicts the hierarchical organization of communities, these methods show low matching percentage scores with PR and also failed to exhibit modular organization and hemispheric symmetry of nodes.

3.6 Conclusions

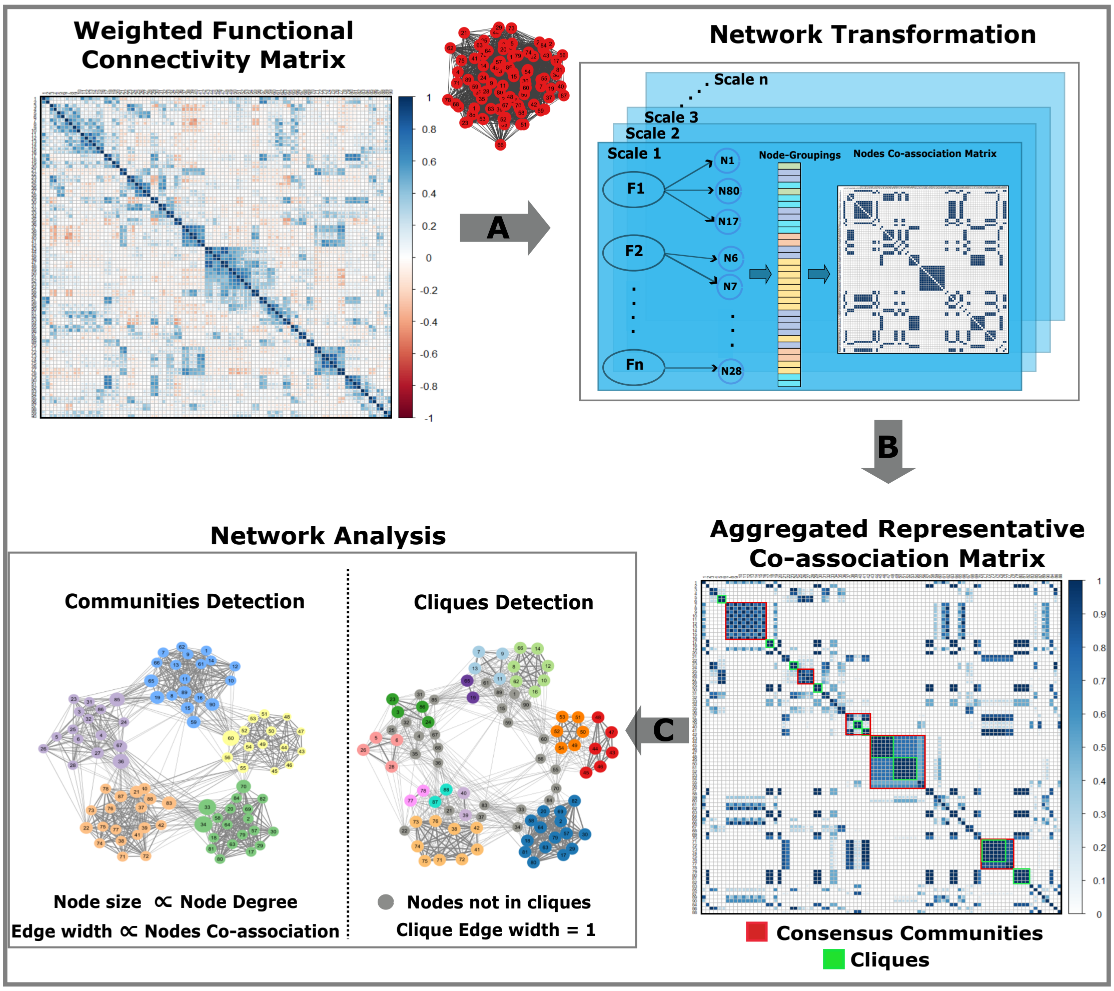
Overall, we conclude that when derived the optimal number of communities of the network, both the methods LM and EFA perform equivalently. In this case, EFA with five factors and LM at $\tau = 0.4$, i.e., with five communities, behaves similarly and complies with PR. Additionally, EFA also exhibits bilateral symmetry, maximized modularity, and hierarchical modular organization. The novelty of our work lies in using EFA for community detection, which is traditionally used for correlation analysis in statistical applications. Here, we avoid sparsification or binarization of the network and use a fully connected network for functional segregation, i.e., identifying functionally significant communities.

Limitations: EFA expects an input parameter, namely, the number of factors, n_F, for its implementation. Here, n_F corresponds to the number of node-partitions in FCN. However, there is no *ground truth* for a definite number of node-partitions in FCN. Moreover, the hierarchical brain organization provides several valid counts, depending on the level we choose. Also, in general, EFA suffers from the issues such as replicability of node-partitioning [39], and the generalizability of the algorithm [40].

Hence to address these limitations of EFA, we define the value of n_F as the *scale*, and propose to apply EFA for multiple scales to get optimal node-groupings [41]. In Chapter 4, we address the limitations of EFA by using a *multiscale consensus method* to find modularity maximized *communities* and tight-bound *cliques* within communities.

CHAPTER 4

MULTISCALE CONSENSUS APPROACH FOR BRAIN FUNCTIONAL SEGREGATION



For the effective use of EFA for FCN analysis, we implement a novel consensus-based algorithm using a multiscale approach (Figure FC1.1), with the number of factors n_F as a scale (Step-**A**). The consensus mechanism is used for transforming the network (Step-**B**), where we perform community detection and cliques on the transformed network (Step-**C**).

In the multiscale consensus approach, we define the value of n_F as the *scale* for EFA, and propose to apply EFA for multiple scales. The node-partitioning is represented using a co-association matrix D^k, for the k^{th} scale. We choose a set of values of n_F to compute EFA, and the communities/node-groupings are aggregated to generate a final co-association matrix, D, using consensus voting of the D^k at $k = 1, 2, \ldots, N$ scales. Thus, we transform a weighted, fully connected FCN (correlation matrix) to a representative co-association matrix. D is symmetric, with the values in the range of 0 to 1, 0 for the nodes that did not group together in any of the node-partitionings obtained in any of the chosen scales, and 1 for nodes that always belonged in the same community through all the node-partitionings. Thus, we use D as an adjacency matrix of our *transformed* network [41]. Generalized Louvain (genLouvain) [1] community detection algorithm is then implemented on the transformed network, to find the consensus-based communities and cliques. Our work is different from the state-of-the-art consensus methods to find the communities in the networks [134–136], owing to the use of EFA.

In this Chapter, we describe the following steps that are used to study the brain FCN:

1. **Network Transformation** (Section 4.1) has three sequential steps.
   (a) *Input data preparation:* This step involves collecting rs-fMRI data, processing data for our case study, aggregating subject-wise correlation matrices, and creating a weighted, undirected, fully connected FCN (Section 4.1.1).
   (b) *Multiscale co-association matrix computation:* We first explain how node-partitioning is done using EFA with a predefined scale, *i.e.*, number of factors

n_F (Section 4.1.2). We then identify different values of n_F for implementing EFA at multiple scales (Section 4.1.3), and implement node-partitioning at each of the N scales, n_F^k, for $k = 1, 2, \ldots, N$. The node-partitioning at the k^{th} scale is represented using a co-association matrix D^k. We, thus, determine N such co-association matrices.

(c) *Transformed network generation:* We aggregate the co-association matrices at multiple scales using consensus voting to generate a representative co-association matrix, which is the adjacency matrix of a transformed network (Section 4.1.4).

2. **Communities and Cliques in Transformed Network:** Finding groups of nodes in the transformed network, in the form of *communities* and *cliques* (Section 4.2).

3. **Optimal Selection of Scales:** Finding appropriate scales for the multiscale implementation is critical in our methodology. Hence, we study how the choice of scales impacts our results by measuring the efficiency of its outcome, *i.e.*, communities. We use appropriate efficiency metrics for deciding the optimal selection of scales, as explained in Section 4.3.

4. **Significance of Consensus Communities and Cliques:** Using the identified optimal scale, we obtain consensus communities, and cliques and study them for their biological significance (Section 4.4).

4.1 Network Transformation

Here, we describe our proposed methodology for network transformation.

4.1.1 Input data preparation and FCN generation

We have run experiments on datasets of FCN of the human brain in a resting state, with different sizes, and using different parcellation atlases (AAL, Schaefer). In Chapter-3, we had performed a single case study on 90-nodes FCN using AAL atlas, which we expand here to include datasets with the state-of-the-art Schaefer parcellation, with 200-nodes and 400-nodes FCNs. We include an FCN with 400 nodes in our case study to explore the cortical brain areas at high resolution, as explained in their work by Essen et al. [137] and Schaefer et al. [59]. Hence, we have used two different fMRI datasets. The first dataset gives $\mathbf{SCH}_{400}$ and $\mathbf{SCH}_{200}$, which are networks using parcellations determined by Schaefer *et al.* [59], and the second one, $\mathbf{AAL}_{90}$, which is a network with AAL parcellation atlas [58]. This selection of FCNs enables us to study the influence of both 'network size' and 'parcellation atlas' in our results.

Schaefer's parcellation dataset gives an eyes-open (EO) resting-state functional connectome from the enhanced Nathan Kline Institute-Rockland Sample (NKI-RS)[1] [138]. This data is available for open access using the Amazon S3 web services bucket. Multiband and multiplexed echo-planar imaging (EPI) [139, 140] is used to acquire fMRI data. The datasets were preprocessed to correct for slice timing, nonlinear distortion, and motion using fMRIPrep v1.1.8 [141]. The data was bandpass filtered (0.008-0.08 Hz), linearly detrended, and nuisance regressed using a 36+ parameter strategy [142]. A preprocessed time series for each subject is generated by averaging the data within each node (ROI) and fitted to each subject's anatomy [143]. This data is of 109 subjects (56 female and 53 male) of the age group, 18-26 years, and the FCNs confirm to 17 reference networks [6]. The same AAL parcellation data used in Chapter-3 (Section 3.2) is used in this Chapter on the multiscale consensus workflow. All three datasets, i.e., $\mathbf{SCH}_{400}$, $\mathbf{SCH}_{200}$, and $\mathbf{AAL}_{90}$ are from resting-state fMRI modality, of healthy right-

[1] `http://fcon_1000.projects.nitrc.org/indi/enhanced/`

handed controls, with the right gender-balance and strictly following the age group of 18-26 years. The studies [6, 144, 145] show decreased segregation and modularity with age. Hence, considering young adult subjects for FCN functional segregation is very necessary to learn the patterns in the brain.

Functional Connectivity Network Generation: The FCN from all the datasets of our interest are generated using mean time courses (BOLD signal) of the nodes in the ROIs given by the corresponding parcellation technique. Pairwise, Pearson's correlation among all the nodes gives an undirected, weighted, and completely connected FCN. We generate a final correlation matrix by aggregating individual FCN matrices, which is an adjacency matrix. $\mathbf{SCH}_{200}$ is an adjacency matrix of size 200, $\mathbf{SCH}_{400}$ is of size 400, and $\mathbf{AAL}_{90}$ is of size 90.

4.1.2 EFA and Co-association Matrix

We *partition* the FCN by identifying factors corresponding to node-partitions in the network. The complete details of factor analysis are presented in Chapter-3, Section 3.4.1. The four-step algorithm described in Chapter-3 is upgraded to a five-step model to apply EFA on FCN using a multiscale approach: (1) checking the feasibility of implementing EFA on the corresponding correlation matrix, (2) estimating n_F, (3) implementing EFA using MLE with both rotation methods, (4) retaining node-grouping from communality-maximizing rotation method, and (5) generating a co-association matrix from the node-grouping.

As a first step to implementing EFA, the *eligibility criterion* is verified. All three correlation matrices, i.e., $\mathbf{SCH}_{400}$, $\mathbf{SCH}_{200}$, and $\mathbf{AAL}_{90}$ are, positive semidefinite and are with the measure of sampling adequacy (MSA) of 0.97, 0.96, and 0.78, respectively. Thus, these FCNs are qualified for node-partitioning using EFA. We follow the same FRC as explained in Chapter-3, i.e., to use only the factors which have at least one node correlated with the factor and to ensure the inclusivity of all nodes in the FCN in the node-partitions, we considered all factor loading with $f_s > 0$. Steps 1-4 are the same as explained in Chapter-3 (Section 3.4.1).

Co-association Matrix: The node-partitioning is represented as a co-association matrix, which is used for consensus gathering. The co-association matrix is symmetric with the FCN nodes in rows and columns in the same permutation order, and the matrix element is a binary truth value for the nodes in the corresponding row and column belonging to the same partition, P. Thus, the co-association matrix for a scale, n_F^k, where P_i^k and P_j^k are the node-partitions/factors of EFA, containing nodes i and j is:

$$D_{ij}^k = \begin{cases} 1 & \text{if } P_i^k = P_j^k, \\ 0 & \text{otherwise.} \end{cases} \tag{Eqn 4.1}$$

4.1.3 Multiscale EFA

We implement EFA at multiple scales to generate different sets of node-partitions, and then take a consensus of their corresponding co-association matrices to determine the functional segregation of the dataset. Our goal is to find the most efficient set of n_F as scales, thus improving the generalizability and replicability of the results. We measure the efficiency using metrics explained in Section 4.3. In order to find a set of scales n_F, we use a continuous interval of integral values and, as a trivial choice, a singleton, *i.e.*, a single scale.

4.1.3.1 Continuous Interval for n_F:

Finding a continuous interval of n_F is a trivial way of finding multiple scales for our approach. This interval I, which consists of only positive integers, is expected to be a property of the FCN. I consists of all the possible values of the number of node-partitions, n_P, that the FCN can have. In our work, we have considered n_F consistently from (a) a priori knowledge (n_F^{pr}), (b) using parallel analysis (n_F^{pa}), (c) scree test (n_F^{st}), and (d) using graph-theoretic community detection methods Louvain [126] (n_F^{lc}), and Infomap [146] (n_F^{im}). The FCN is a completely connected network, and it is a requirement not to filter the network to apply EFA. However, finding n_F^{lc} and n_F^{im} using community detection methods require filtering the network to reduce the edge density in FCN. We use percolation analysis [115] to find an appropriate threshold value to filter and sparsify the networks and then apply Louvain and Infomap community detection methods to identify n_F^{lc} and n_F^{im}. Note that Louvain and Infomap community detection methods are used exclusively to find the number of communities to provide values for determining the bounds of the interval for n_F. Hence, the outcomes of these methods are not used subsequently in determining the final consensus n_P.

Overall, we use n_F^{pr}, n_F^{pa}, n_F^{st}, n_F^{lc} and n_F^{im} to determine the lower and higher bound values for the continuous interval for n_F. We have proposed this approach for finding multiple scales in our work [41].

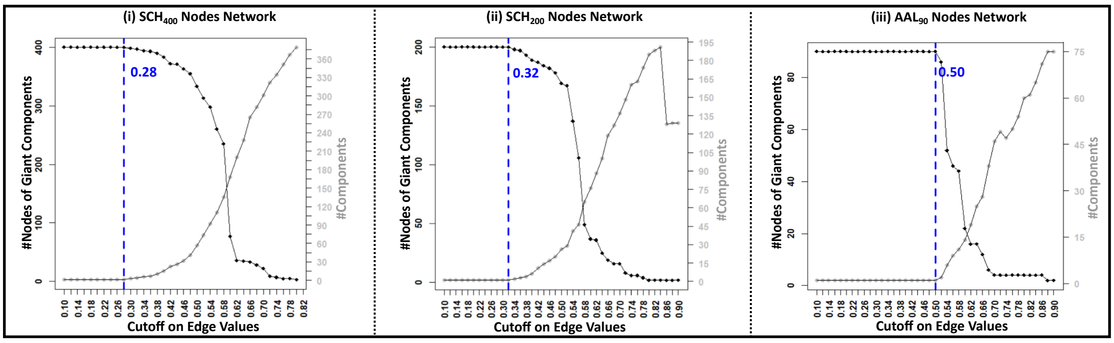


Figure FC4.1: Finding edge cutoff value to sparsify the network using percolation analysis, (i) **SCH**$_{400}$ nodes network, the cutoff value is *0.28* (ii) **SCH**$_{400}$ nodes network, the cutoff value is *0.32* (iii) **AAL**$_{90}$ nodes network, the cutoff value is *0.5*.

To implement Louvain and Infomap community detection methods to sparsify the FCN, we have used percolation analysis. The threshold for network sparsification is selected by verifying the number of nodes of the largest connected components as we remove edges in decreasing order of edge-weights using percolation analysis. In percolation analysis, an edge-weight value is considered as a threshold, when at values higher than the edge-weight, the network tends to break down into a larger number of smaller subnetworks. Our results show that the optimal threshold for the **SCH**$_{400}$ network is **0.28** that retains 7360 edges; for the **SCH**$_{200}$ network, it is **0.32**, retaining 1901 edges; and for the **AAL**$_{90}$ network, it is **0.5**, retaining 231 edges (Figure FC4.1). That implies edge filtering retains only 4.6%, 4.7%, and 2.9% of the edges in the completely connected network in **SCH**$_{400}$, **SCH**$_{200}$, and **AAL**$_{90}$, respectively. As shown in Figure FC4.1, at the cutoff values below the optimal one, the network stays as a single connected component, and beyond the optimal value, the giant connected components (GCC's) of the network start disintegrating into the sub and sub-sub components, and the number of connected components increases.

Table TC4.1: Multiscale interval selection using values for n_F from different sources for the selected datasets.

	n_F^{pr}	$n_F^{\text{pa}}, n_F^{\text{st}}$	n_F^{lc}	n_F^{im}	Selected Interval I
SCH$_{400}$	{**7**, 17} [6]	**19**	6	10	**[7, 19]**
SCH$_{200}$	{**7**, **17**} [6]	13	5	11	**[7, 17]**
AAL$_{90}$	{**5**} [4, 147]	9	7	**12**	**[5, 12]**

To determine the continuous interval I for finding multiple scales, we have collected the n_F. Table TC4.1 lists these n_F values for our selected datasets, where a list of values for n_F^{pr} is included. Both parallel analysis and scree test have given the same values for each of the selected datasets. Using the extrema from these values, we determine the multiscale interval I for **SCH**$_{400}$, **SCH**$_{200}$, and **AAL**$_{90}$ to be **[7, 19]**, **[7, 17]**, and **[5, 12]**, respectively.

4.1.3.2 Goodness of fit for EFA

We first check the goodness of fit for factors found in each integer values in I by examining the root mean square residual (RMSR) and the root mean square error of approximation (RMSEA) values for the correlation matrices for our selected FCNs.

Table TC4.2: Root mean square residual (RMSR) and root mean square error of approximation (RMSEA) values for n_F values in the interval I selected for the datasets.

#Factors	**SCH$_{400}$**		**SCH$_{200}$**		**AAL$_{90}$**	
	RMSR	RMSEA	RMSR	RMSEA	RMSR	RMSEA
5	NA	NA	NA	NA	0.09	0.04
6	NA	NA	NA	NA	0.07	0.03
7	0.05	0.00	0.05	0.00	0.06	0.00
8	0.05	0.00	0.05	0.00	0.05	0.00
9	0.04	0.00	0.04	0.00	0.05	0.00
10	0.04	0.00	0.04	0.00	0.04	0.00
11	0.04	0.00	0.04	0.00	0.04	0.00
12	0.03	0.00	0.03	0.00	0.03	0.00
13	0.03	0.00	0.03	0.00	NA	NA
14	0.03	0.00	0.03	0.00	NA	NA
15	0.03	0.00	0.03	0.00	NA	NA
16	0.03	0.00	0.03	0.00	NA	NA
17	0.02	0.00	0.02	0.00	NA	NA
18	0.02	0.00	NA	NA	NA	NA
19	0.02	0.00	NA	NA	NA	NA

RMSR is the correlation matrix mean residual measurement; the values RMSR$<$ 0.05 or RMSR$\approx$ 0.00 are recommended as then the derived factors describe the correlation structure at low RMSR values. The RMSR values for selected n_F values for our chosen datasets are given in Table TC4.2. The RMSEA$<$ 0.05 is considered to be a model with good fitness [148, 149]. In our experiments, RMSEA has been found to be 0.0 for all selected n_F across all datasets, except for **AAL**$_{90}$, RMSEA$=$ 0.04 for $n_F = 5$, and RMSEA$=$ 0.03 for $n_F = 6$. Since RMSEA values are favorable for all cases of n_F in the datasets, where RMSR values are marginally greater than the threshold 0.05, we continue to use all scales for our experiments; thus, retaining all scales in the selected intervals.

4.1.3.3 Choice of Rotation Method for EFA

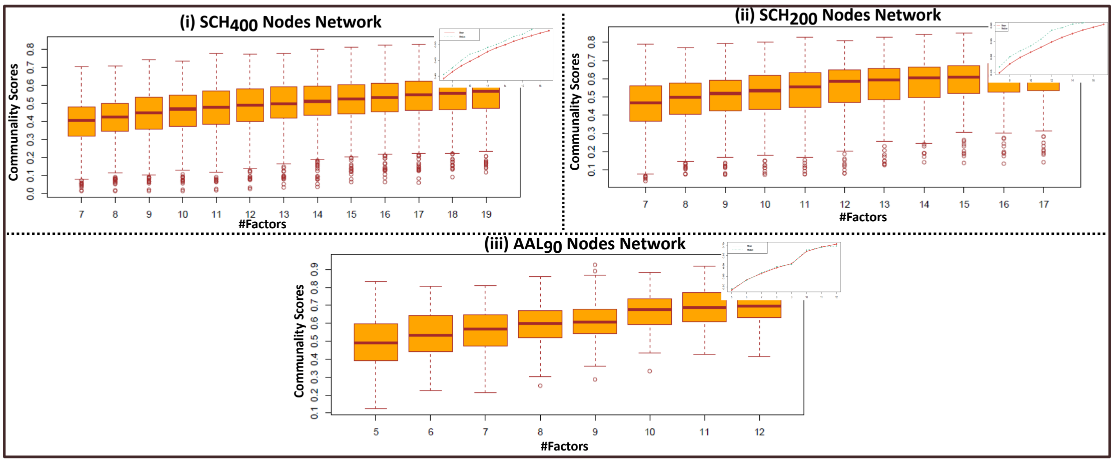


Figure FC4.2: Distribution of communality scores for EFA using maximum likelihood and varimax rotation method at n_F given in the interval I. (i) **SCH**$_{400}$ for I = [7, 19], (ii) **SCH**$_{200}$ for I = [7, 17], and (iii) **AAL**$_{90}$ for I = [5, 12]. Graphs in the inset show the mean and median communality scores for values of $n_F \in I$.

As described in Section 3.4.1 (EFA Implementation Parameters), we have compared the communality scores among the varimax and oblimin rotation methods and used varimax as its merits over oblimin method. For our experiments, Figure FC4.2 shows the distribution of communalities scores at each scale, *i.e.*, n_F of EFA. We observe that for all FCNs, the communality scores are significant, and the median value at each scale is greater than 0.4. We conclude that EFA with the maximum likelihood method and with varimax rotation produces acceptable error terms and communality scores in the correlation matrices for the FCNs.

4.1.3.4 Sub-interval from Ensemble Experiments

Since there are several choices for multiple scales, we run an ensemble of multiscale EFA to identify the optimal set of multiple scales for node-partitioning. We select the optimal set of multiple scales using a data-driven ranking scheme based on the efficiency metrics (Section 4.3). As a first cut, we use sub-intervals in I as a selection of scales to generate the ensemble, while the exhaustive set can include a random selection of scales within I. When we use the continuous interval, $I = [lb,\ ub]$ for different scales to be used for multiscale EFA, with bounds lb and ub, we get n_{SI} valid sub-intervals, such that $n_{SI} = \binom{ub-lb+1}{2}$, using valid combinations of the bounds. For instance, for $\mathbf{SCH}_{200}$, $n_{SI} = \binom{17-7+1}{2} = 55$.

To determine the optimal set of scales, these different sub-intervals allowable within I are run as an ensemble of experiments. We further aggregate D for a set of possible sub-intervals in n_{SI} (Section 4.1.4), for a multiscale approach with consensus voting. We, thus, get node-partitioning or communities for each dataset from the implementation of generalized Louvain community detection on the transformed network, represented by aggregated D, for each experimental run. Since Louvain community detection gives randomized node-partitioning, we run genLouvain method 1000 times for each experimental run. Once we run an ensemble set of experiments, we determine the optimal sub-interval that maximizes the efficiency of the node-partitioning, measured using normalized mutual information (NMI), modularity score (Q), Silhouette score (S), and Dunn index (DI), as explained in Section 4.3. Thus, multiscale implementation with ensemble runs provides the exploratory and experimental characteristics of EFA application in FCN analysis.

4.1.4 Transformed Network Generation

Fred et al. [150–152] have discussed how the individual clustering results of an ensemble are combined to create a co-association matrix by voting the nodes that are grouping together. Similarly, we aggregate the N co-association matrices from running EFA for N different scales by a consensus voting process. The consensus results are known to produce more stable node-partitions/factors/groups/communities [134, 135]. Our voting process aggregates by averaging the co-association matrices from all the scales. Thus, we effectively compute an $[n \times n]$ consensus matrix D, which gives the likelihood of two nodes in the FCN that co-associate across different scales.

$$D_{ij} = \frac{\sum_{k=1}^{N} D_{ij}^{k}}{N} \qquad \text{(Eqn 4.2)}$$

Here, the number of scales, N, is specifically the continuous interval range, e.g., $I = [7,17]$ has 11 scales. D can now be treated as the adjacency matrix of a newly transformed network, N$_T$, of the FCN.

4.2 Communities and Cliques in Transformed Network

We first find communities in N$_T$, and then identify cliques within the communities to analyze the network and further study the biological significance of our consensus communities and cliques.

Consensus Communities of the FCN: We perform a community detection on the representative matrix, i.e., the transformed network, N$_T$, using Generalized Louvain (genLouvain) algorithm. GenLouvain is a variant of the Louvain method [1], which works on a modularity matrix and is used to find final consensus communities. Louvain community detection method [126] has been used widely for finding communities in a network. Louvain is a greedy optimization method that works on adjacency matrix to find communities using modularity measures. The modularity measure with a resolution parameter that is used for optimization in genLouvain method is [153]:

$$Q(\vec{c},\gamma) = \frac{1}{2m} \sum_{i,j=1}^{n} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta\left(c_i, c_j\right), \qquad \text{(Eqn 4.3)}$$

where A is an adjacency matrix, γ is a resolution parameter that plays a major role in finding communities, and $k_i = \sum_j A_{ij}$, and $2m = \sum_i k_i$.

The consensus matrix conventionally densifies a sparse network [134]; however, in our work, D sparsifies a dense (full) network, *i.e.*, FCN.

Consensus Cliques of the FCN: After binarizing the consensus communities in the transformed networks, we get connected components, which are subnetworks with edge weight $D_{ij} = 1$. We observe that these connected components are cliques, the FCN which correlate strongly to the same factor demonstrate *cliqueness*, which is a significant result of our work.

Alternative to the method of finding cliques within the communities, we can find them directly as connected components in the edge-filtered transformed network, where we retain only edges with an edge weight 1.0.

We observe that the cliques identified in our work could not have been found directly from the FCN as the correlations between nodes in the cliques are not necessarily strong. Each clique, thus, is a combination of strongly as well as weakly correlated node-pairs. Our methodology shows that these weak edges are significant, as they are part of cliques. Thus, the conventional procedure of thresholding/filtering an FCN would discard such cliques in the community as weak correlation edges get filtered out.

4.3 Efficiency Metrics to Find Optimal Sub-interval Selection

We have used an ensemble using different possible sub-intervals of the selected continuous interval, I, for a set of multiple scales. We use the following four measures to evaluate the partitioning from each experimental run: normalized mutual information (NMI), modularity score (Q), Silhouette score (S), and Dunn index (DI). The choice of these metrics is based on their performance in measuring the efficiency of node-partitioning. We expect more efficient node-partitioning to have a higher modularity score, and confirm with the prior knowledge, i.e., 5 modules of [4] for $\mathbf{AAL}_{90}$, and 17 networks of [6] for **SCH**$_{400}$, and **SCH**$_{200}$ datasets.

4.3.1 Efficiency Metrics

We define our selected efficiency metrics here.

Normalized Mutual Information (NMI): This is a widely used measure to compare the node-partitions. It is a ratio of mutual information [154] between the partitions to the individual entropy [155] values. The NMI, as defined by Fred and Jain [136] is:

$$NMI\left(\vec{C_1},\vec{C_2}\right) = \frac{2 \cdot I\left(\vec{C_1},\vec{C_2}\right)}{H\left(\vec{C_1}\right) + H\left(\vec{C_2}\right)}, \qquad \text{(Eqn 4.4)}$$

where $I(\vec{C_1},\vec{C_2})$ is the mutual information between the communities $\vec{C_1}$, and $\vec{C_2}$, and $H(\vec{\cdot})$ is the entropy of each community. As this is a normalized measure, we can compare the node-partitions with a different number of communities.

Modularity (Q): Measures the node-partitions score using their intra-community edge-density and the expected edge-density in a random network. The detailed description of Q is given in Section 3.5 (Eqn 3.1, Eqn 3.2).

Silhouette Score (S): Silhouette score is a measure for evaluation of clusters without ground truth. The higher the Silhouette score better the defined clusters of the data. The score is measured as [156]:

$$s = \frac{(b-a)}{\max(a,b)}, \qquad \text{(Eqn 4.5)}$$

where a, b is the distance measure of intra- and inter-cluster edges, respectively. We have used a dissimilarity matrix to measure S, the value close to 1 defines a better cluster organization of nodes.

Dunn index (DI): Dunn index is widely used to quantify the clusters in the absence of ground truth. *DI* is a ratio of the smallest inter-cluster distance to the largest intra-cluster distance [157]. *DI* is computed on a dissimilarity matrix, similar to **S**.

4.3.2 Optimal Sub-interval of Multiple Scales

We visualize the efficiency metrics for the ensemble of experiments using a dot matrix. Each dot represents an experiment based on a specific sub-interval of scales. The plot is implemented as an upper-left triangle to reflect the choice of lower and upper bounds for each experiment. Thus, the diagonal reflects the single-scale implementation. The results of efficiency metrics for the different datasets are juxtaposed in Figure FC4.3, for identifying the experiment with optimal sub-interval for multiple scales. We observe that multiple experiments give optimal results across the four metrics.

To quantitatively determine the most optimal sub-interval, we normalize the four efficiency metrics, and score each experiment using the average, maximum, and median of its normalized values. The choices of the statistical function to use for ranking have been based on the scores we have obtained for the datasets of our interest. Hence, we refer to this process as a *data-driven ranking* of the ensemble results. Here, we rank all the experiments for a dataset based on the average, maximum, and median scores. We then take the sum of ranks to identify the top-ranked experiment for each dataset. We get the following optimal sub-interval [11, 18] for $\mathbf{SCH}_{400}$, [7, 10] for $\mathbf{SCH}_{200}$, and [5, 6] for $\mathbf{AAL}_{90}$. However, for $\mathbf{AAL}_{90}$, we observe that the selected scales have higher RMSR (Table TC4.2) and RMSEA. Hence, we choose the next optimal sub-interval, namely [5,11], as the most optimal sub-interval for $\mathbf{AAL}_{90}$.

From Figure FC4.3, we can reconfirm that sub-intervals **[11, 18]**, **[7, 10]**, and **[5, 11]** give the sets of multiple scales for $\mathbf{SCH}_{400}$, $\mathbf{SCH}_{200}$, and $\mathbf{AAL}_{90}$, respectively, for which all the efficiency metrics are maximized, thus, giving the most optimal scale selection for finding communities and cliques. Using these selected sub-intervals, we

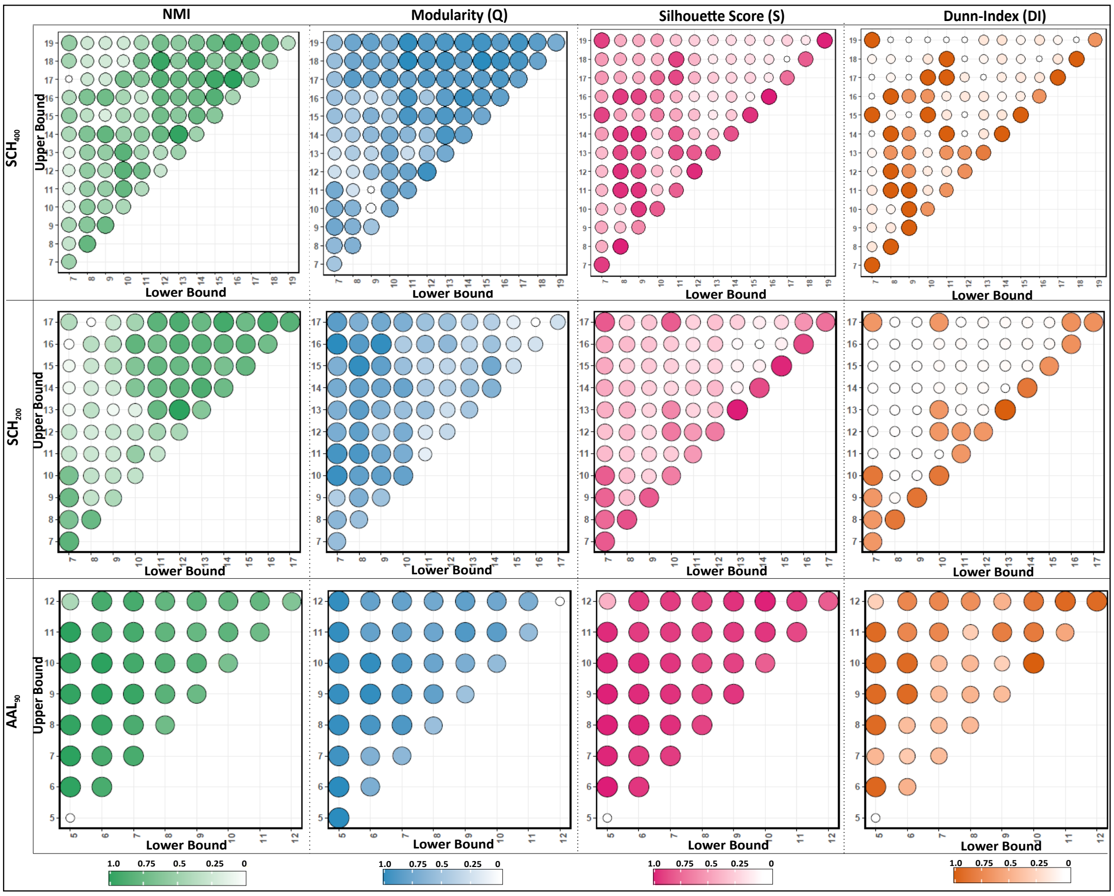


Figure FC4.3: The dot matrix is used for visually comparing the efficiency metrics of the communities or node-partitioning obtained in each experimental run of the ensemble for each dataset. The diagonal corresponds to the single-scale EFA, and the upper left-triangle format indicates the choice of lower and upper bounds of the sub-interval *I* used for each experiment. Both the size of the circle glyph and its color visually encode the value of the efficiency metric. For finding an optimal outcome of consensus communities, we identify a sub-interval that maximizes all measures, i.e., NMI, Q, S, and DI. For **SCH**$_{400}$, **SCH**$_{200}$, and **AAL**$_{90}$, **[11, 18]**, **[7, 10]**, and **[5, 11]** are chosen as the optimal sub-intervals that maximizes all the metrics, respectively.

demonstrate our final results of the multiscale consensus communities and cliques, which are computed as explained in Section 4.2.

4.4 Multiscale Consensus Communities and Cliques

The consensus communities and cliques are the results of our proposed methodology. Here, we report specific characteristics of our results.

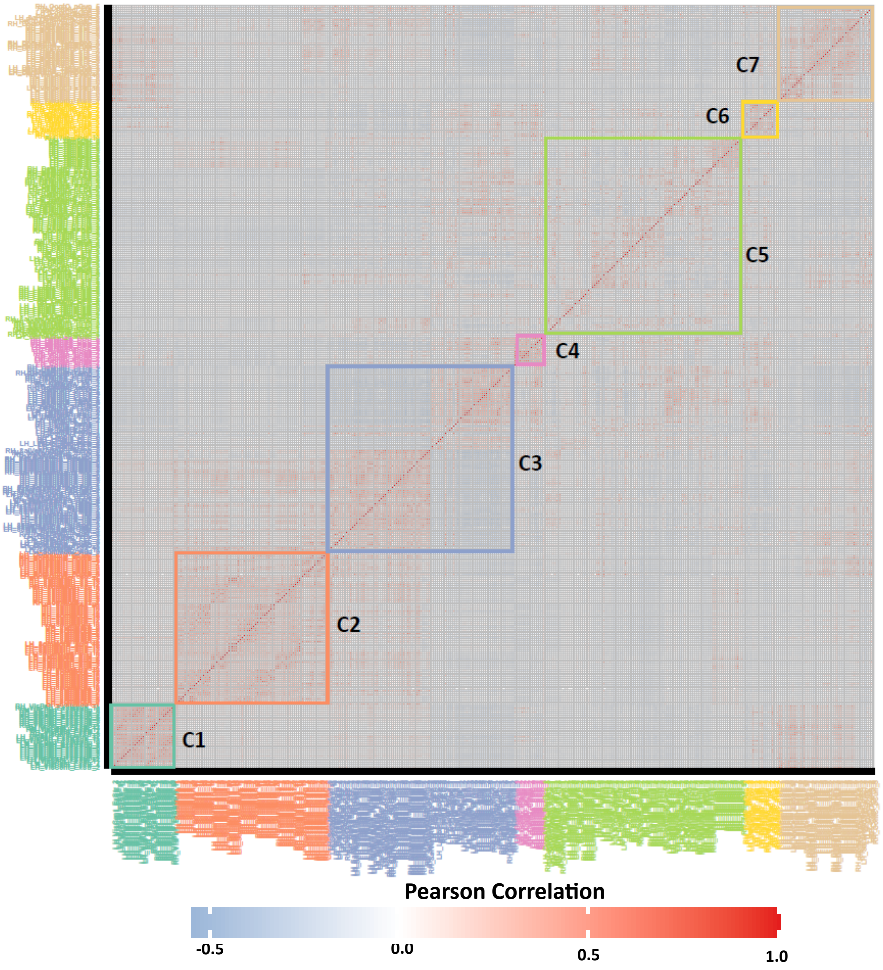


Figure FC4.4: Matrix visualization of consensus communities from multiscale EFA of **SCH**400 network, identified using optimal sub-interval [11, 18], that gives seven consensus communities, where nodes within communities are seriated based on brain regions are given by Schaefer parcellation (Table TC4.3).

4.4.1 Consensus Communities

The multiscale consensus communities in $\mathbf{SCH}_{400}$ and $\mathbf{SCH}_{200}$ are visualized using seriated matrices in Figures FC4.4 and FC4.5, respectively, and those of $\mathbf{AAL}_{90}$ are shown on the brain surface with an axial view in Figure FC4.6. We use IDs prefixed with '**C**' and '**Cq**' for communities and cliques, respectively, hereafter.

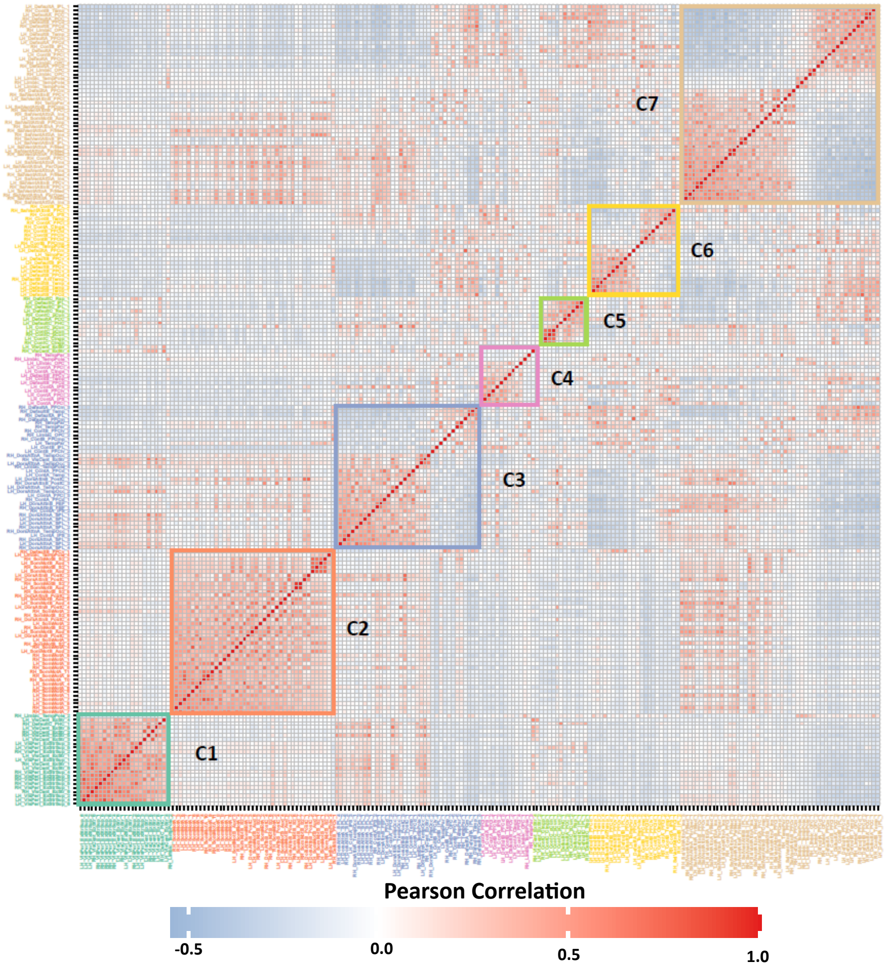


Figure FC4.5: Matrix visualization of consensus communities from multiscale EFA of $\mathbf{SCH}_{200}$ network, identified using optimal sub-interval [7, 10], that gives seven consensus communities, where nodes within communities are seriated based on Principal Component Analysis.

The matrix visualization requires seriation, *i.e.*, reordering, of nodes to reveal block-like structures along the diagonal, which are patterns for clusters. There are several methods to do seriation [158]. We first seriate the nodes based on the community ID

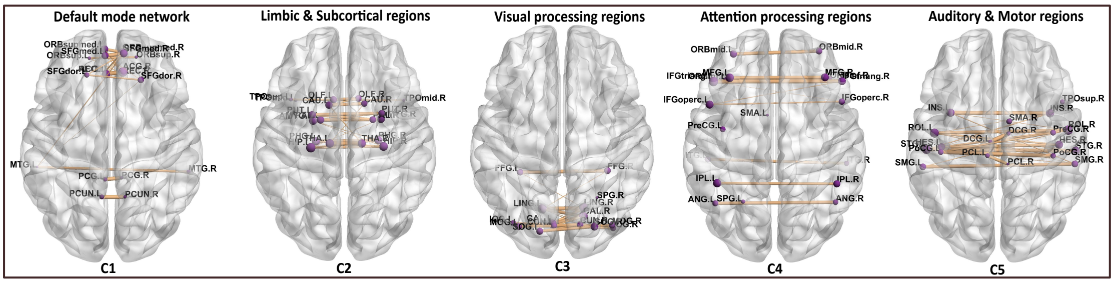


Figure FC4.6: Communities from multiscale EFA of **AAL**$_{90}$ network, identified using the optimal sub-interval [5, 11], that gives five consensus communities.

from our results. We choose a seriation method for seriating nodes within a community that reveals the node-partitioning as block structures along the diagonal. We thus use the ordering using principal component analysis (PCA) for **SCH**$_{200}$ (Figure FC4.5), and lexicographical ordering based on the brain regions in Schaefer parcellation [6] for **SCH**$_{400}$ (Figure FC4.4).

Table TC4.3 shows how well the multiscale consensus communities in **SCH**$_{400}$ and **SCH**$_{200}$ match with the different regions of the brain identified as per the naming convention used in Schaefer parcellation [6], the same which has been used for lexicographical ordering for matrix seriation. We observe that each consensus community predominantly contains large subnetworks belonging to multiple regions in Schaefer parcellation. While in **SCH**$_{200}$ FCN, each community has subnetworks belonging to at most two regions, we observe that in **SCH**$_{400}$ FCN, the fragmentation is more pronounced and each community consists of subnetworks belonging to at most four regions. This clearly shows that our consensus communities have communities within themselves, indicating the future scope for finding hierarchical communities.

For **AAL**$_{90}$ FCN, we compare our results with a similar dataset investigated by He *et al.* [4] in Figure FC4.7. We use a novel summary visualization using a square to indicate a node, five parallel axes to represent the five communities in Figure FC4.7 (i), a horizontal line representing the line of symmetry between the two lateral hemispheres in Figure FC4.7 (ii), and in both, we use the color of the square to indicate

Table TC4.3: The matching of our best results of communities (*Cmm.*), each with N_c nodes, extracted using multiscale EFA with the eight reference subnetworks (*Ref.Subnet.*) [6] given as number of matching nodes. The reference subnetworks, each with N_{sn} nodes, are obtained from the naming convention of corresponding regions in Schaefer parcellation, namely *Occipital Lobe* (**OL**), *Default Mode Network* (**DMN**), *Executive Control Network* (**ECN**), *Dorsal Attention* (**DA**), *Limbic System* (**LS**), *Somato Cortex* (**SC**), *Salience Attention* (**SA**), *Temporal Lobe* (**TL**).

Ref.Subnet. → Cmm. ↓	**OL**	**DMN**	**ECN**	**DA**	**LS**	**SC**	**SA**	**TL**
(i) **SCH**$_{400}$, using optimal sub-interval [11, 18]								
N_{sn}	47	79	61	52	24	70	51	16
C1 N_c=33	**33**	0	0	0	0	0	0	0
C2 N_c=80	0	0	0	12	1	**67**	0	0
C3 N_c=98	0	**30**	7	9	7	3	**42**	0
C4 N_c=15	0	**7**	**8**	0	0	0	0	0
C5 N_c=105	0	**25**	**38**	2	**16**	0	8	**16**
C6 N_c=19	0	**17**	2	0	0	0	0	0
C7 N_c=50	**14**	0	6	**29**	0	0	1	0
(ii) **SCH**$_{200}$, using optimal sub-interval [7, 10]								
N_{sn}	24	37	37	22	14	34	26	6
C1 N_c=23	**21**	1	0	0	1	0	0	0
C2 N_c=41	0	1	0	6	1	**33**	0	0
C3 N_c=36	1	4	**10**	**16**	2	0	0	3
C4 N_c=13	0	2	**8**	0	2	0	0	1
C5 N_c=14	2	**6**	**6**	0	0	0	0	0
C6 N_c=23	0	**11**	**9**	0	1	0	1	1
C7 N_c=50	0	**12**	4	0	7	1	**25**	1

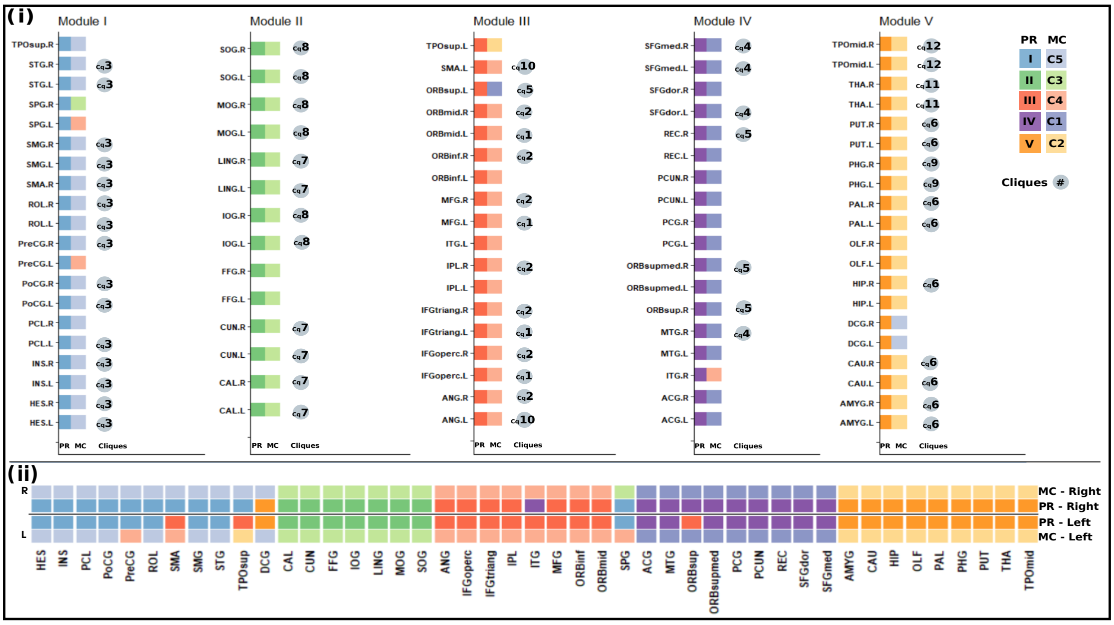


Figure FC4.7: An abstract summary visualization of comparing our results of five consensus communities from multiscale EFA with optimal sub-interval [5,11], with the modules I-V from a similarly parcellated dataset [4]. (i) Multiscale consensus (MC) communities in comparison with modules I-V from prior results (PR), and also indicating the nodes in MC which are found to be clique nodes. (ii) Identifying the extent of bilateral symmetry of nodes from both MC and PR communities, where L and R correspond to the left and right hemispheres of the brain.

the community to which the node belongs. The darker shades indicate communities in prior results (PR) [4], and the lighter shades, our results using multiscale consensus communities (MC). We observe that only eight out of 90 nodes do not match, *e.g.*, bilateral SPG (superior parietal gyrus) in Module-1 of PR and bilateral DCG (Median cingulate and paracingulate gyri) of Module-V are four nodes that do not match with our results. Figure FC4.7 (i) and (ii) show the extent of the match in community composition and bilateral symmetry, respectively. We observe that the bilateral symmetry is marginally higher in our results than in the prior results. Thus, our visualization considers non-spatial and spatial matching of the results. The community C3 maps exactly with Module-II in Figure FC4.7 (i), which corresponds to the occipital region. C3 is thus consistent with prior work [4, 159]. Overall, we conclude that our results of the communities bear closeness to published results in the relevant literature.

4.4.2 Consensus Cliques

We observe that as we increase the number of scales in the multiscale implementation, the cliques get more *compact*. This is because, as we gather consensus from more scales, the edges with weight $D_{ij} = 1$ get sparser, thus generating more *compact* cliques. Table TC4.4 gives the salient differences in the cliques we find when using the optimal sub-interval and the entire continuous interval. Since cliques are tightly and completely connected subnetwork, using the latter is applicable here. Table TC4.5 maps out the cliques present in each of the multiscale consensus communities in our selected datasets. We observe that 63%, 55%, and 70% of the nodes in FCN of **SCH**$_{400}$, **SCH**$_{200}$, and **AAL**$_{90}$, respectively, are cliques, thus indicating a large enough subnetwork in the FCN being cliquish. The number of nodes in cliques in a community is roughly proportional to the community's size. At the same time, large communities have at most two large cliques and several fragmented cliques.

Table TC4.4: Comparison of the cliques obtained when using the optimal sub-interval and the entire continuous interval.

SCH$_{400}$	
Optimal sub-interval [11, 18] – 305 nodes in 42 cliques of sizes {36, 36, 30, 21, 20, 16, 15, 12, 11, 8, 6, 6, 6, 5, 5, 5, 4, 4, 4, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2}	Entire interval [7, 19] – 252 nodes in 44 cliques of sizes {36, 34, 30, 16, 14, 12,12, 7, 5, 5, 4, 4, 4, 4, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2}
SCH$_{200}$	
Optimal sub-interval [7, 10] – 162 nodes in 30 cliques of sizes {28, 23, 20, 13, 8, 6, 6, 5, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2}	Entire interval [7, 17] – 110 nodes in 20 cliques of sizes {21, 17, 17, 11, 6, 5, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2}
AAL$_{90}$	
Optimal sub-interval [5, 11] – 66 nodes in 11 cliques of sizes {16, 13, 7, 6, 6, 4, 4, 4, 2, 2, 2}	Entire interval [5, 12] – 63 nodes in 12 cliques of sizes {15, 9, 7, 6, 6, 4, 4, 4, 2, 2, 2, 2}

The multiscale consensus cliques on brain surface with axial view are shown for the three datasets in Figure FC4.8. Additionally, the matrix visualization of the completely connected subnetworks in the cliques also give an understanding of fine-grained tightly-packed modular organization of the brain, as shown for **AAL**$_{90}$ in Figure FC4.9 (i).

4.4.3 Significance of Resultant Communities and Cliques

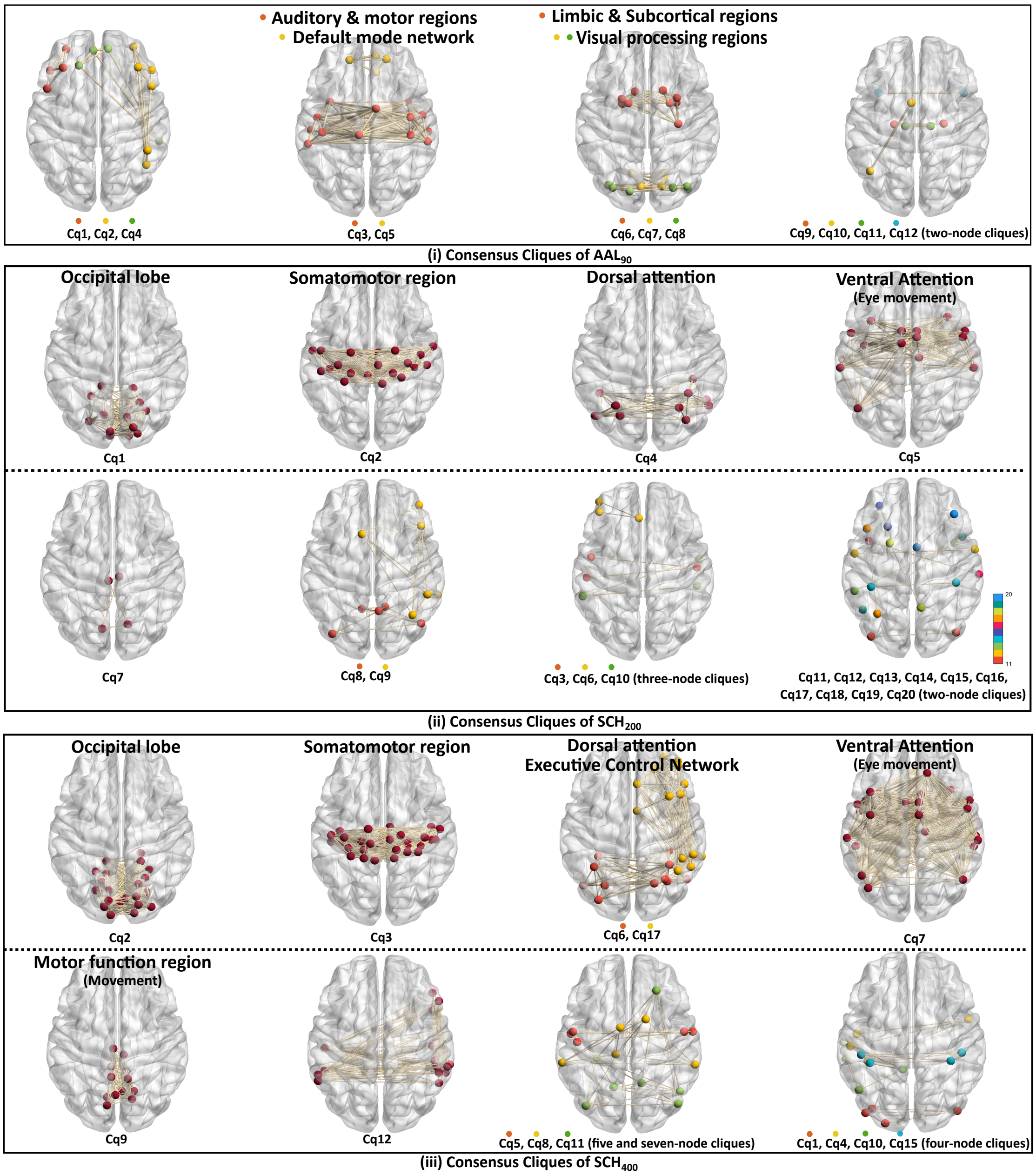
As discussed earlier, the resultant communities signify a group of subnetworks from multiple specific, reference functional regions of the brain. This also indicates the scope for another level of nested communities to separate the region-wise subnetworks from our communities. We observe that as the network size reduces, our communities contain subnetworks from a lesser number of functional regions. We observe that the communities and cliques found in **SCH**$_{400}$ and **SCH**$_{200}$ have significant correspondences, indicating the consistency of the outcomes of our algorithm.

To quantify the significance of the cliques, we compute the score of each clique. We first consider factor loading (FL) values of each node pertaining to the respective clique. In EFA, factor loadings are the correlations of the random variables with the factors. We now compute the ratio of minimum to maximum FL values for each clique at each scale of the multiscale EFA, and average the scores across the scales of the interval (I) to find the final score of the cliques, thus giving a score, CqS. The CqS values are bounded in [0,1]. The smaller the difference between FL min and max values, the greater the CqS, thus, signifying all of the variables in that factor are correlated with similar FL values. We have observed that most often, a two-node clique exhibited a CqS$>=$ 0.9. For example, for **SCH**$_{200}$, the CqS$>$ 0.9 for cliques Cq11, Cq12, Cq13, Cq14, and Cq19.

AAL$_{90}$ nodes network resulted in 12 cliques with 63 nodes (Figures FC4.8 (i), FC4.9 (i)). Cq9, Cq11, and Cq12 have some of the highest clique scores CqS, and consist of subnetworks from the bilateral parahippocampus, thalamus, and middle-temporal gyrus, respectively. The stimuli such as an image of the landscape make the parahip-

Table TC4.5: Cliques within multiscale consensus communities ($\mathrm{C_m}$), with N_c nodes and N_{cq} cliques. Clique ID's are given in italic font, and the number of nodes of the respective clique is given in parenthesis; for example, *Cq2*(30) reads as *Clique2* with 30 nodes.

$\mathrm{C_m}$	Set of Cliques – {ID (# Nodes)}		
	SCH$_{400}$ $N_{cq} = 44$ with 252 nodes	**SCH**$_{200}$ $N_{cq} = 20$ with 110 nodes	**AAL**$_{90}$ $N_{cq} = 12$ with 63 nodes
C1	$N_c = 33, N_{cq} = 1$ with 30 nodes {*Cq2*(30)}	$N_c = 23, N_{cq} = 2$ with 19 nodes {*Cq1*(17), *Cq11*(2)}	$N_c = 18, N_{cq} = 2$ with 8 nodes {*Cq4*(4), *Cq5*(4)}
C2	$N_c = 80, N_{cq} = 9$ with 60 nodes {*Cq3*(36), *Cq4*(4), *Cq5*(5), *Cq13*(3), *Cq15*(4), *Cq20*(2), *34*(2), *35*(2), *38*(2)}	$N_c = 41, N_{cq} = 5$ with 30 nodes {*Cq2*(21), *Cq3*(3), *Cq12*(2), *Cq14*(2), *Cq16*(2)}	$N_c = 19, N_{cq} = 4$ with 15 nodes {*Cq6*(9), *Cq9*(2), *Cq11*(2), *Cq12*(2)}
C3	$N_c = 98, N_{cq} = 9$ with 57 nodes {*Cq7*(34), *Cq8*(5), *Cq10*(4), *Cq16*(3), *Cq18*(3), *Cq26*(2), *Cq28*(2), *Cq30*(2), *Cq39*(2)}	$N_c = 36, N_{cq} = 3$ with 16 nodes {*Cq4*(11), *Cq10*(3), *Cq18*(2)}	$N_c = 15, N_{cq} = 2$ with 12 nodes {*Cq7*(6), *Cq8*(6)}
C4	$N_c = 15, N_{cq} = 1$ with 12 nodes {*Cq9*(12)},	$N_c = 13, N_{cq} = 3$ with 7 nodes {*Cq6*(3), *Cq15*(2), *Cq17*(2)}	$N_c = 19, N_{cq} = 3$ with 13 nodes {*Cq1*(4), *Cq2*(7), *Cq10*(2)}
C5	$N_c = 105, N_{cq} = 11$ with 49 nodes {*Cq12*(16), *Cq17*(14), *Cq19*(3), *Cq23*(2), *Cq24*(2), *Cq25*(2), *Cq27*(2), *Cq29*(2), *Cq40*(2), *Cq41*(2), *Cq44*(2)}	$N_c = 14, N_{cq} = 2$ with 9 nodes {*Cq7*(4), *Cq8*(5)}	$N_c = 19, N_{cq} = 1$ with 15 nodes {*Cq3*(15)}
C6	$N_c = 19, N_{cq} = 3$ with 11 nodes {*Cq11*(7), *Cq42*(2), *Cq43*(2)}	$N_c = 23, N_{cq} = 3$ with 10 nodes {*Cq9*(6), *Cq19*(2), *Cq20*(2)}	
C7	$N_c = 50, N_{cq} = 10$ with 33 nodes {*Cq1*(4), *Cq6*(12), *Cq14*(3), *Cq21*(2), *Cq22*(2), *Cq31*(2), *Cq32*(2), *Cq33*(2), *Cq36*(2), *Cq37*(2)}	$N_c = 50, N_{cq} = 2$ with 19 nodes {*Cq5*(17), *Cq13*(2)}	


● Auditory & motor regions
● Default mode network
● Limbic & Subcortical regions
● ● Visual processing regions
Cq1, Cq2, Cq4
Cq3, Cq5
Cq6, Cq7, Cq8
Cq9, Cq10, Cq11, Cq12 (two-node cliques)
Occipital lobe
Somatomotor region
Dorsal attention
Ventral Attention
(Eye movement)
Cq1
Cq2
Cq4
Cq5
Cq7
Cq8, Cq9
Cq3, Cq6, Cq10 (three-node cliques)
Cq11, Cq12, Cq13, Cq14, Cq15, Cq16,
Cq17, Cq18, Cq19, Cq20 (two-node cliques)
20
11
Occipital lobe
Somatomotor region
Dorsal attention
Executive Control Network
Ventral Attention
(Eye movement)
Cq2
Cq3
Cq6, Cq17
Cq7
Motor function region
(Movement)
Cq9
Cq12
Cq5, Cq8, Cq11 (five and seven-node cliques)
Cq1, Cq4, Cq10, Cq15 (four-node cliques)

(i) Consensus Cliques of AAL$_{90}$

(ii) Consensus Cliques of SCH$_{200}$

(iii) Consensus Cliques of SCH$_{400}$

Figure FC4.8: Visualization of the consensus cliques on the brain surface (axial view) (i) **AAL**$_{90}$ network with 63 nodes in 12 cliques using interval [5, 12] for multiscale EFA, (ii) **SCH**$_{200}$ network with 110 nodes in 20 cliques using interval [7, 17] for multiscale EFA, and (iii) **SCH**$_{400}$ network with 252 nodes in 44 cliques using interval [7, 19] for multiscale EFA, of which (4+)-node cliques are shown here. Predominant regions are labeled here.

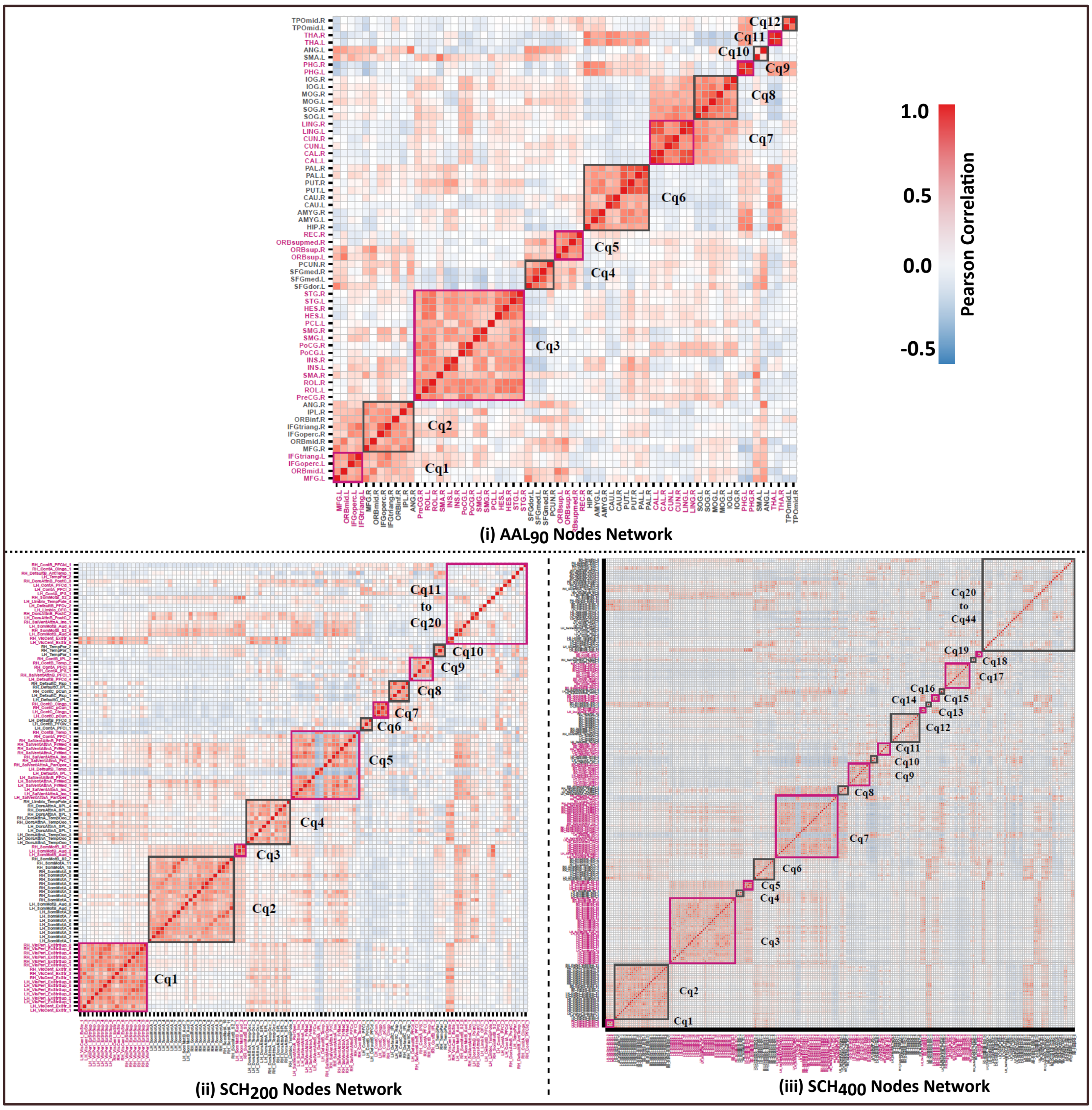


Figure FC4.9: Network visualization using seriated matrices of (i) 12 cliques with 63 nodes of **AAL**$_{90}$ using the interval [5, 12], (ii) 20 cliques with 110 nodes of **SCH**$_{200}$ using the interval [7, 17], and (iii) 44 cliques with 252 nodes of **SCH**$_{400}$ using the interval [7, 19]; for multiscale EFA.

pocampus region highly active, but we notice this region is a clique in our study of rs-fMRI data. This indicates active visual responses even when the subjects are in an eyes-closed state while fMRI scans. All nodes of the thalamus featuring as a clique is interesting, as this region serves as a process and relay station to sensory systems. The middle-temporal gyrus is in the temporal lobe of the brain, and is known for visual perception, speech, and semantic memory processing. All 12 nodes of the occipital lobe that are the reason for vision and perception are included in Cq7 with bilateral calcarine, cuneus, and lingual gyrus nodes, and Cq8 with bilateral superior, middle, and inferior bilateral nodes. In the eyes-closed state, the occipital lobe regions across the subjects express similar patterns, and hence they are observed as closely-knit networks in Cq7 and Cq8.

SCH$_{200}$ network has 20 cliques with 110 nodes (Figures FC4.8 (ii), FC4.9 (ii)). Cq11 is with bilateral VisCent_ExStr_5 nodes of the occipital region, and Cq14 is with bilateral DorsAttnB_PostC_2 nodes of dorsal attention. Cq12, both the nodes belong to the somatosensory cortex region, Cq13, and Cq19 though displayed high clique-scores; each node belongs to different brain regions. Interestingly, though the clique-score is < 0.9, in Cq1, Cq2, Cq3, Cq7, and Cq10, all nodes of these cliques belong to the occipital lobe with bilateral symmetry, somatomotor cortex, somatosensory cortex, executive control network, and temporal lobe, respectively. In Cq4, out of 11 nodes, 10 nodes are from dorsal attention regions; as our study is on resting-state scans, maybe self-directed thoughts and awareness around the surroundings are significant among the subjects, causing these regions to form a clique. **SCH**$_{400}$ nodes network resulted in 44 cliques with 252 nodes (Figures FC4.8 (iii), FC4.9 (iii)) of which the cliques with the scores CqS$>=$ 0.9 are of total nine, two-nodes cliques. Three of these cliques are from the default mode network, two are from the bilateral occipital lobe, two cliques belong to the bilateral somato-cortex, one clique is of the bilateral temporal region, and one clique is from the executive control network. As the dataset used for generating

$\mathbf{SCH}_{400}$ and $\mathbf{SCH}_{200}$ nodes network is the same, we can observe similar clique patterns. The Cq1 of $\mathbf{SCH}_{200}$ (Figures FC4.8 (ii)) and Cq2 of $\mathbf{SCH}_{400}$ (Figures FC4.8 (iii)) are exclusive of occipital lobe nodes, similar to Cq7 and Cq8 of $\mathbf{AAL}_{90}$ (Figure FC4.8 (i)). Similarly, Cq2 and Cq3 of $\mathbf{SCH}_{200}$ and $\mathbf{SCH}_{400}$ networks are the bilateral somato cortex nodes, Cq4 and Cq6 of $\mathbf{SCH}_{200}$ and $\mathbf{SCH}_{400}$ networks are the nodes of bilateral dorsal attention.

Overall, the cliques from our algorithm show significant links or edges in the FCNs, which cannot be otherwise identified using the correlation values. Figure FC4.9 shows that cliques make significant block patterns along the diagonal, indicating the dominant presence of high correlation values. At the same time, we observe that the edges do not necessarily always correspond to high correlation values, *e.g.*, Cq11 of $\mathbf{AAL}_{90}$. Interestingly, edges of negative correlations are also captured in cliques, *e.g.*, Cq5 and Cq7 of $\mathbf{SCH}_{400}$ and $\mathbf{SCH}_{400}$ networks. The conventional edge filtering process used for functional segregation fails to preserve these structures, and hence would not have detected these cliques. Our cliques serve as a *compact cover* of significant edges for functional segregation, where the significance is by virtue of co-association and not just the correlation values.

4.5 Conclusions

Our motivation is to exploit the semantics of the correlation matrix to avoid the edge filtering step, for finding salient node partitioning of the network. Hence, we use EFA, which is orchestrated by using a consensus method, where EFA is run with multiple values of the number of factors n_F, used as a scale. Thus, we propose an algorithm to find consensus communities and cliques using the EFA method with multiple scales for various parcellation dimensions of fMRI data. The correctness of EFA is verified using the RMSR, and RMSEA values and also noticed significant communality scores for varying scales of EFA. Finding the range interval and identifying the sub-interval to find final communities adds to the exploratory nature of the factor analysis. We have identified modularly organized communities in all three case studies, visualized the same using matrix visualization with seriation and also on brain surface axial view using spatial centroid coordinates. Interestingly, the cliques found in all case studies are the closely-knit subnetworks that are observed across all scales; these ROIs always grouped and formed cliques irrespective of the change of scale. The identified communities and cliques are studied for their biological significance and compared with the relevant prior studies. Owing to the smaller size of the network, the bilateral symmetry of ROIs and comparison of consensus communities have been demonstrated for $\mathbf{AAL}_{90}$, and the results are comparable with the prior studies. Exploring EFA with a range of scales and identifying the appropriate sub-interval produces a maximized modularly organized brain regions, i.e., communities. Our proposed algorithm is thus scalable to the size of the FCN, and is generalizable to different parcellations used in constructing the FCN. Overall, our method shows how a conventional correlation analysis, namely EFA, can be effectively used with network-based approaches for functional brain studies.

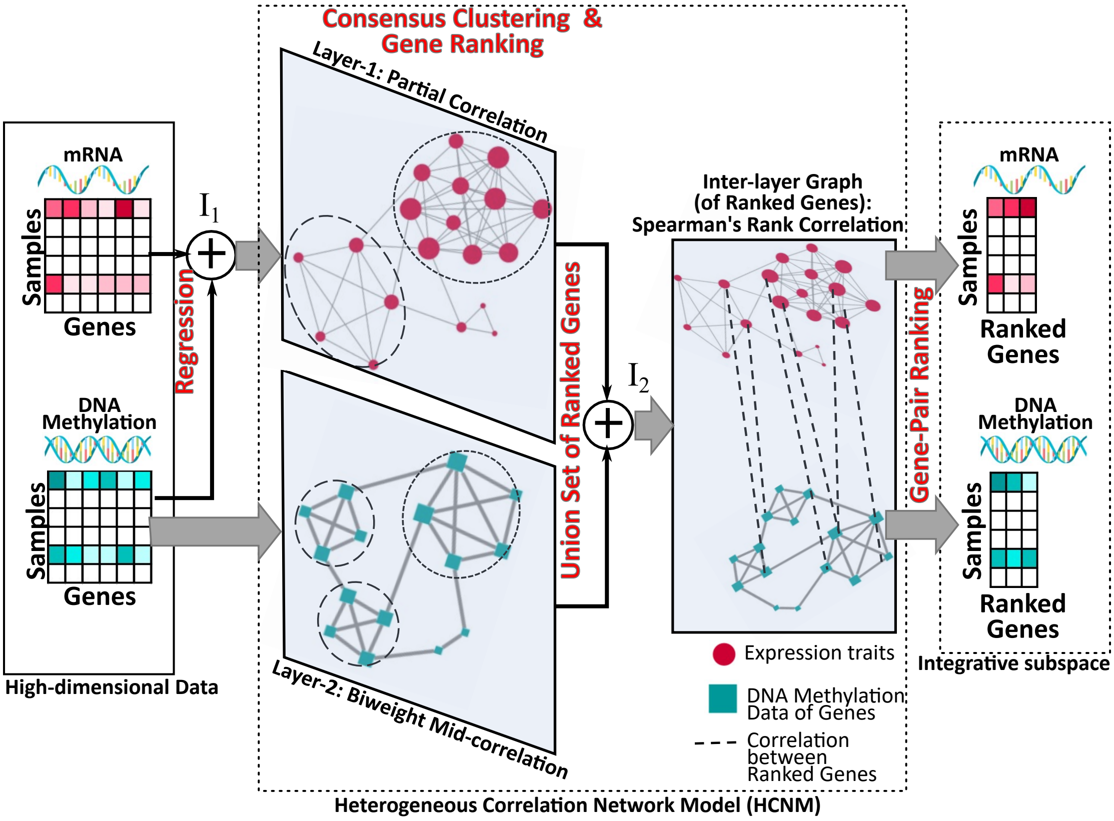
Limitations: Since our algorithm uses EFA as a central step, there is a strict requirement of positive-definiteness for the correlation matrix of the FCN. While a correlation matrix by definition must be positive-definite, it is not guaranteed owing to the complex preprocessing methods implemented for converting fMRI data to the correlation matrix and aggregating matrices across subjects. It is yet to be studied how a correlation matrix additionally corrected to be positive definite would work with our algorithm.

Part II

Multi-level Integrative Study of Multi-omics Cancer Data

CHAPTER 5

REPRESENTATIVE INTEGRATIVE SUBSPACE OF MULTI-OMICS


Consensus Clustering & Gene Ranking
Layer-1: Partial Correlation
Layer-2: Biweight Mid-correlation
mRNA
Samples
Genes
DNA Methylation
Samples
Genes
High-dimensional Data
I_1
Regression
Union Set of Ranked Genes
I_2
Inter-layer Graph (of Ranked Genes): Spearman's Rank Correlation
Gene-Pair Ranking
● Expression traits
■ DNA Methylation Data of Genes
- - - Correlation between Ranked Genes
Heterogeneous Correlation Network Model (HCNM)
mRNA
Samples
Ranked Genes
DNA Methylation
Samples
Ranked Genes
Integrative subspace

There have been recent efforts in comprehensive studies of “multidimensional” omics data [46], which in oncology has been encouraged by the release of The Cancer Genomic Atlas (TCGA) dataset [160]. TCGA provides genomic, epigenomic, transcriptomic, and proteomic data of various cancer profiles, facilitating researchers to study significant cancer-causing genes and cancer subtypes using both single- and multi-omic features. These comprehensive studies are conducted by *integrating* either the data, its analytics, or both from these different omic features [46].

For cancer studies pertaining to outcome prediction, multi-omics information has been routinely integrated at the data-level to obtain transformed data models, such as, regression and network models. For instance, multivariate multiple linear regression of multi-omics data has been used to construct gene-gene interaction (GGI) networks [42], and directed random walks with multi-omic information has been used on pathway information [44]. Recently, the multi-omics information has been integrated to form a discriminative dimensionality reduction tree [43], which is further used for outcome prediction. Dimensionality reduction of omic features is generally mandated owing to their unbalanced dimensionality, *i.e.*, fewer samples and many more omic features [43, 45]. The available high-throughput omic data causes a “small n, large p” or “short-fat data” problem. The network topology-based algorithms can alleviate this problem through its gene ranking applications. Identifying these significant genes and using them as representative features creates “low-dimensional subspaces” [161]. The representative subspace is the subspace that best represents the full space for subtype classification.

In this work, we use methylation features and expression traits of *breast* and *lung* cancer profiles of TCGA database. We apply our proposed multi-level integrative workflow to both the phenotypes (Figure FC1.2). Each of the state-of-the-art integrative studies has its own benefits and shortcomings, and are mostly used in isolation. We also observe that the integrative studies broadly fall under the category of data modeling or transformation. We hypothesize that the semantics of some of these data models allows them to be extendable, and also work with other integrative methods. We consider a specific example of extending the use of an integrative regression model for finding representative subspaces, followed by an appropriate network fusion method for predicting cancer subtypes. The integrative regression model captures the interdependence between two multi-omics features at the data-level [42], whereas the network fusion integrates the analytics performed separately from different omic features (Chapter 6). Thus, we demonstrate that such integrative methods can be plugged into the same workflow or implementation to improve the overall understanding of the high-dimensional multi-omics data. In order to achieve a multi-level integration of the multi-omics data through existing integrative methods, namely regression, and network fusion, we propose a data model that will transition one method to another, referred to as the *Heterogeneous Correlation Network Model* (HCNM). We propose a three-level integration algorithm driven by HCNM for gene-ranking, integrative subspace identification, and cancer subtype prediction (Figure FC1.1). Finding integrative subspaces is equivalent to feature selection as well as dimensionality reduction, and is a pertinent research problem in the face of increased dimensionality in integrative studies [46].

Heterogeneous networks are a special class of multilayer networks, where the nodes in each layer are different [46]. Heterogeneous networks have intra-layer and inter-layer graphs, where the latter is a bipartite graph between nodes in different layers [47]. These networks are predisposed to embed the multi-omics data by design, and thus provide novel tools for integrative studies [46]. Our proposed data model is specifically a heterogeneous *correlation* network model. HCNM is similar to the heterogeneous network model, iHNMMO [162] in terms of the use of regression and correlation. The difference is that iHNMMO has normalized correlation network layers with regression coefficients as inter-layer graph edge weights, whereas HCNM has a partial correlation layer, which is an intra-layer, computed from the regression model, and cross-correlation coefficients as inter-layer graph edge weights (Figure FC1.2).

Here, we propose a multi-step algorithm for the construction and use of HCNM to find representative integrative subspace of multi-omics. The steps are: (1) integration using multivariate multiple linear regression I_1, (2) construction of correlation network layers for intra-layer graphs, (3) community detection by consensus, (4) ranking genes to be integrated in an inter-layer graph I_2, (5) computing inter-layer graph edge weights, thus completing our HCNM, and (6) finding integrative subspace by ranking edges in inter-layer graph.

The novelty of HCNM lies in embedding the interdependence of different omic features in intra-layer edges, rather than inter-layer edges. Our contributions are in:

- Transforming a network-free multivariate multiple linear regression model to our proposed heterogeneous correlation network model, HCNM,
- Using consensus voting in the intra-layer graphs of HCNM for ranking genes,
- Proposing an algorithm with multi-level integration of multi-omics data for gene-subspace identification, and cancer subtype identification (Chapter 6).

The following sections describe the details of the data set and preprocessing procedure (Section 5.1), followed by our multi-step process for finding representative integrative subspace (Section 5.2), and the significance of the genes of identified subspace (Section 5.3).

5.1 Data and Preprocessing

Our case study pertains to ‘breast invasive carcinoma’ (TCGA-BRCA) and ‘lung squamous cell carcinoma’ (TCGA-LUSC) of the TCGA database. We have used an R, Bioconductor package *TCGAbiolinks* [163] to download mRNA expression data from *Illumina HiSeq platform* and DNA methylation data from the *Illumina Human Methylation 450 platform*. Along with omic data, the clinical data of all samples of each cancer profile are collected from the TCGA database[1].

[1] The dataset has been downloaded in December 2020.

The three broad steps followed for preprocessing the data are a) outlier removal, b) imputing missing values, and c) standardization of the data. We perform outlier removal for each omic dataset by removing the features that satisfy one of these three conditions: (1) its value across all samples is zero, (2) its missing values account for more than 25% of the overall sample size, (3) its variance is in the lower 25% of the overall variance of all features [100, 164]. For the retained omic features, we impute the missing values using the median value of all samples. For each omic feature, we then standardize the values using z-scores, such that $(\mu, \sigma) = (0, 1)$. Finally, methylation probes are mapped to genes; and if a probe is mapped to multiple genes, a least correlated feature with the gene expression trait is considered [165]. Suppose the gene is not available in expression data and multiple probes are associated with it. In that case, the methylation feature with the maximum variance is considered and mapped to that gene.

Table TC5.1: The dataset dimension details before and after preprocessing.

Dataset		Breast Cancer	Lung Cancer
Samples	From TCGA	1,098	504
	Preprocessing	718	319
	Final Data	***486***	***200***
mRNA	From TCGA	19,947	19,947
	Preprocessing	16,626	16,877
	Final Data	***16,626***	***16,877***
DNA Methylation	From TCGA	445,577	485,577
	Preprocessing-Step1	395,669	395,958
	Preprocessing-Step2	54,868	45,665
	Preprocessing-Step3	41,190	32,703
	Final Data	***10,109***	***9,405***

The samples of each cancer data are first filtered using clinical data. The omic features samples are those with the tumor sample-type that is either ‘primary tumor’ or ‘metastatic’; the cases with the status, ‘solid tissue normal’ are not considered in this study. The samples are filtered out if a patient’s vital status is ‘alive’, yet the survival days are less than the median of all subject’s overall survival days. The final selected samples/patients are common subjects of both the omic features. The mRNA data is filtered by excluding the genes with more than 25% of zeros in expression values. methylation features are first filtered based on missing values, i.e., methylation probes with more than 25% of NAs or missing values are removed (Preprocessing-Step1, Table TC5.1). The remaining NAs and missing values are replaced with the median of the respective patient data. The methylation probes are further filtered if the variance is less than 25% of the maximum variance of all probes (Preprocessing-Step2, Table TC5.1). Finally, methylation probes are mapped to genes; and if a probe is mapped to multiple genes, a least correlated gene with the expression trait is considered. Suppose the gene is not available in expression data and multiple probes are associated with it. In that case, the methylation feature with the maximum variance is considered and mapped to that gene (Preprocessing-Step3, Table TC5.1). The final data dimensions of breast cancer data are 486 samples, 16626 expression traits, 10109 methylation features, and lung cancer data are with 200 samples, 16877 expression traits, and 9405 methylation features. Both mRNA and DNA methylation features data are normalized, such that each gene has zero mean and unit standard deviation.

5.2 Finding Integrative Subspace Using HCNM

Here, we construct the intra-layer graphs in HCNM, detect communities by consensus in these layers, rank genes based on communities to construct the inter-layer graph, and finally construct the inter-layer graph in HCNM. Using the inter-layer graph, we perform a second iteration of ranking genes, to select highly ranked "significant" genes for finding a representative subspace. We refer to this as an *integrative* subspace as it is the union-set of subspaces in all omic feature spaces, which includes two integration steps: I_1 using regression to construct one of the intra-layer graphs, and I_2 where genes are selected using consensus communities and ranking procedure.

Step-1: Multivariate Multiple Regression (I_1*):*

We use the MMR model by treating DNA methylation data as features and expression traits as outputs, for integrating selected methylation features and expression [42].

For m mRNA expression values and n methylation features, we have $Y \in \mathbb{R}^{k\times m}$ and $X \in \mathbb{R}^{k\times n}$, respectively, for k samples. The MMR model is written as

$$Y = X \cdot B + E, \ where\ B \in \mathbb{R}^{m\times n}\ and\ E \in \mathbb{R}^{m\times k}, \qquad \text{(Eqn 5.1)}$$

for the regression coefficient matrix B and residual error matrix E. We now have $Y = [y_1, y_2, \ldots, y_m]$, corresponding to $E = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m]$, with $\varepsilon_i \sim N(0, \sigma^2)$, $\forall i \in [1, m]$, by the conventional linear regression model. Here, we implement MMR using Lasso (Least absolute shrinkage and selection operator) [166] regression model.

Step-2: Construction of Intra-layer Graphs in `HCNM`:

To transform a regression model to a network-based model, correlation networks are a natural choice. Regression models have been used for computing partial correlation coefficients [167], which quantifies the correlation between the dependent variables, when conditioning on the independent variables. The intra-layer graph for the independent variable is computed using conventional correlation values.

***Layer-1** (expression traits)*:- When a linear regression model is used, the n$^{\text{th}}$-order partial correlation, *i.e.*, conditioned to n independent variables, can be computed as the total linear (Pearson) correlation between the residual errors [167]. When Y is regressed on X, the residual error $e^{(Y)}$ represents the parts of Y that are uncorrelated with X.

$$e^{(Y)} = Y - \left(\hat{\beta}_0^{(Y)} + X\hat{\beta}_1^{(Y)}\right). \qquad \text{(Eqn 5.2)}$$

Thus, the partial correlation coefficient z of Y, when conditioning on X, is:

$$\mathrm{z}\{Y\} = \{\rho(\varepsilon_i, \varepsilon_j)\} = \rho\left\{e^{(Y)}\right\}, \ where\ \ \rho_{\varepsilon_i,\varepsilon_j} = \frac{\mathrm{cov}(\varepsilon_i, \varepsilon_j)}{\sigma_{\varepsilon_i}\sigma_{\varepsilon_j}}, \qquad \text{(Eqn 5.3)}$$

and cov and σ refer to covariance and standard deviation, respectively. These computed partial correlation coefficients are now weights of edges between m expression traits, in **Layer-1** of `HCNM`.

Layer-2 *(DNA methylation data of genes)*:- Since we are computing the linear correlation amongst the methylation features, we determine the biweight midcorrelation (bicor) coefficients [90]. Bicor is widely used for computing correlation between genomic features, as it is a median-based measure, making it less prone to outliers. Despite their similarities, bicor is preferred over Pearson correlation in genomic applications, where it is also widely used as a similarity measure.

For the methylation feature vectors of samples ‘s’, $m = (m_1, m_2, ...m_s)$ and $n = (n_1, n_2, ...n_s)$, the u_i and v_i for all $i = (1, 2, ...s)$ are defined as [168]:

$$\mathrm{u_i} = \frac{m_i - \mathrm{med}(m)}{K * \mathrm{MAD}(m)}$$

$$\mathrm{v_i} = \frac{n_i - \mathrm{med}(n)}{K * \mathrm{MAD}(n)} \qquad \text{(Eqn 5.4)}$$

Where *med ($\vec{.}$)* is a median of the vector and *MAD ($\vec{.}$)* is the median absolute deviation. Lax [169] had used a measure named *triefficiency* to compare the methods, Wilcox [168] used the same measure and empirically found *K = 9*. In this work, we have used the same value for K. The weights w_i for $\vec{m}$ and $\vec{n}$ are defined as:

$$w_i^{(m)} = \left(1 - u_i^2\right)^2 I\left(1 - |u_i|\right)$$

$$w_i^{(n)} = \left(1 - v_i^2\right)^2 I\left(1 - |v_i|\right) \qquad \text{(Eqn 5.5)}$$

The indicator function $I(1 - |u_i|)$ is 1 if $1 - |u_i| > 0$, and for all other cases indicator function is 0, for the given m and n observations of samples (s), the bi-correlation is given as

$$\tilde{m}_i = \frac{(m_i - \text{med}(m))w_i^{(m)}}{\sqrt{\sum_{j=1}^{s}\left[\left(m_j - \text{med}(m)\right)w_j^{(m)}\right]^2}}$$

$$\tilde{n}_i = \frac{(n_i - \text{med}(n))w_i^{(n)}}{\sqrt{\sum_{j=1}^{s}\left[\left(n_j - \text{med}(n)\right)w_j^{(n)}\right]^2}} \qquad \text{(Eqn 5.6)}$$

$$\text{bicor } (m,n) = \sum_{i=1}^{s} \tilde{m}_i\tilde{n}_i \qquad \text{(Eqn 5.7)}$$

These computed bicor coefficients are now weights of edges between n methylation features, in **Layer-2** of `HCNM`.

Step-3: Community Detection by Consensus:

Clusters of genes in GGI networks, identified using their coexpression values, are often enriched with similar functional annotations [170]. Communities identified in these networks give such gene clusters. Both **Layer-1** and **Layer-2** are completely connected networks, which requires them to be sparsified for performing community detection using popularly used methods, such as, walktrap [171], fast greedy optimization [172], and Louvain community detection [126].

Graph Sparsification:- Wolfe *et al.* [173] have explained how the guilt-by-association (GBA) heuristic implies that the weaker co-expressions or edges in a network are frequently connected to the dissimilar functional clusters. Thus, these edges have to be filtered out from the connected components of the network. These connected components are also *"locally dense, globally sparse"* communities with strong inter- and weak intra-community links. There exist several thresholding techniques for filtering the edges of a graph [174], such as maximal clique, spectral clustering, p-value based cutoff, high pass filter procedure, top 1% of correlations, percolation analysis (PA), etc.

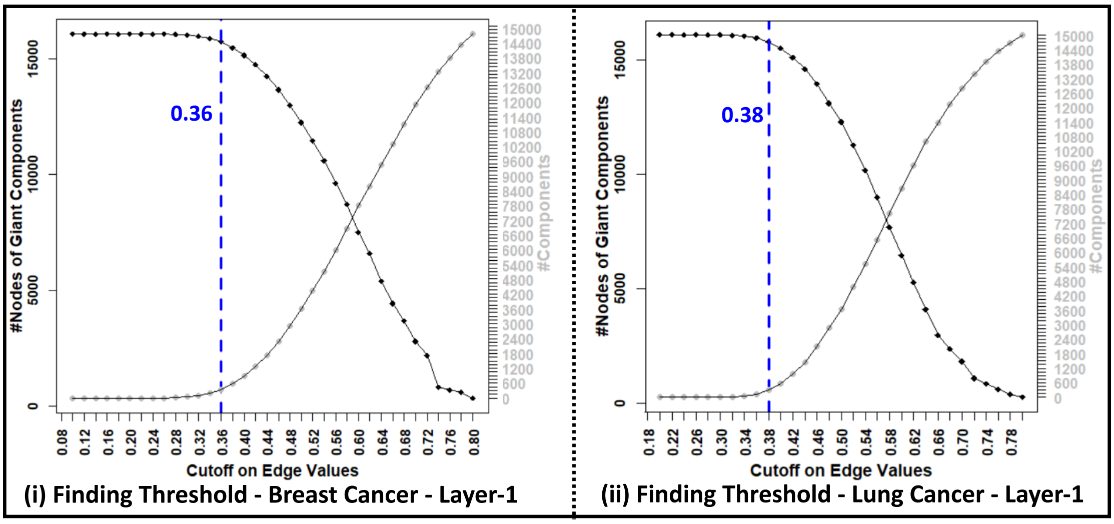


Figure FC5.1: A plot of the edge cutoff value against the network components using percolation analysis, giving selected threshold for intra-layer graph, **Layer-1**, (i) $\tau = 0.36$, for breast cancer, and (ii) $\tau = 0.38$, for lung cancer.

PA involves observing the connected components of the network while progressively increasing the threshold for edge weights [115]. The threshold at which the giant connected component begins to fragment is considered optimal to filter out edges that retain the network as a single connected component. We use this threshold value τ as an absolute value cutoff, implying filtering out edges with weights in the interval $[-\tau, \tau]$. Hence, we first filter the edges based on p-value, and then based on τ from PA. When using correlation networks, we retain only statistically significant edges, which represent correlations with p-value$<$ 0.05. Here, τ for **Layer-1** and **Layer-2** of breast cancer data are 0.36 and 0.32, respectively, and τ of lung cancer networks for **Layer-1** and **Layer-2** are 0.38 and 0.36, respectively (Figure FC5.1). The count of genes after sparsifying the network using PA is given in the Table TC5.2 (Column *Step-3*).

Communities:- "Locally dense, globally sparse" communities regularly occur in biological networks, where hubs distributed in the dense subnetworks play specific biological roles [175]. Depending on the semantics of community formation, intra-community genes have a higher likelihood of similar roles in a specific disease [104]. In the absence of ground truth for communities in the intra-layer graphs in our case study, we find communities by consensus from selected widely used community detection techniques, namely:

Walktrap Method: This method adopts the principle of ‘random walk’ [171]. It exploits the idea that a random walker tends to walk/stay in a dense neighborhood for a longer time than in a sparse one. Identifying such a dense neighbourhood is the key to community detection. The pairwise distance between nodes is computed as the likelihood of reaching from one node to another in n steps, and communities are formed by merging clusters using Ward’s hierarchical clustering [176], while minimizing intra-cluster distances. We use an optimal value of $n = 4$ for our implementation.

Fast Greedy Optimization Method: This method works similar to walktrap using a hierarchical, agglomerative model. It is also similar to Newman’s modularity maximization method [177], but it is more suitable for large networks owing to its linear running time. The greedy optimization helps to find communities with high intra- and low inter-cluster edge densities.

Louvain Community Detection: Louvain is another fast, greedy optimization method that uses modularity maximization to partition a network. The method follows two phases: In the first phase, each node of the network is treated as a community, i.e., the number of nodes in a network is equal to the number of communities. The node N_1 is assigned to its neighbour communities/nodes, for example, N_2, N_5, $N_8, etc.$, and modularity (Q) (Eqn 4.3) is measured; finally, node N_1 is grouped with the node that resulted in maximum modularity score. The process is repeated iteratively until no further increase in modularity score is observed. In the second phase, each community of the first phase is treated as a node where intra-links are nothing, but self-loops and inter-links are represented as weighted edges between the communities. This is widely used in genomic analysis.

Consensus Voting:- We have selected these methods based on the similarity of the semantics of their outputs by design. Thus, we expect to get similar results from these methods, which can be aggregated for a final outcome by consensus. We arrive at a consensus by voting if pairwise nodes, *i.e.*, genes, are likely to be in a community. Thus, the *co-association* votes are equivalent to the likelihood of genes i and j are in the same community across the results from the selected methods. If C_i, C_j are the communities to which genes i, j belong in the method k, the co-association vote is:

$$D_{ij}^{(k)} = \begin{cases} 1 & , \text{if } C_i = C_j, \\ 0 & , \text{otherwise} \end{cases} . \qquad \text{(Eqn 5.8)}$$

The aggregated co-association vote is the average across N_m community detection methods.

$$D_{ij} = \frac{1}{N_m} . \sum_{k=1}^{N_m} D_{ij}^{(k)} . \qquad \text{(Eqn 5.9)}$$

Thus, the network represented by the co-association matrix, D, is the sparsified version of the correlation network, for each intra-layer graph. We now find communities in this transformed network using the genLouvain community detection algorithm [178]. GenLouvain is a variant of the Louvain community detection algorithm that additionally uses a resolution parameter γ to detect communities using a modified modularity score (Eqn 4.3).

Table TC5.2: The total genes at Step-3 after sparsification using PA, Step-4 after consensus-voting and ranking the genes, and Step-6 after gene-pair ranking using inter-layer graph that produces representative integrative subspace.

Subspace in Genes → Omic Features ↓	**Step-3**	**Step-4**	**Step-6**
(i) Breast Cancer			
mRNA	15756	3895	**531**
DNA methylation	9826	1882	**339**
(ii) Lung Cancer			
mRNA	15786	3403	**904**
DNA methylation	9290	3061	**785**

Table TC5.3: Total number of communities identified in **Layer-1** and **Layer-2**.

Commu. Detection → Network Layer ↓	**Walktrap**	**Fast Greedy**	**Louvain**	**Consensus**
(i) Breast Cancer				
Layer-1	2902	626	430	*777*
Layer-2	757	458	310	*130*
(ii) Lung Cancer				
Layer-1	1690	481	333	*1303*
Layer-2	748	397	248	*170*

The number of communities found in each of the layers is given in Table TC5.3. Note that the communities given in Table TC5.3 are without considering the single-node/singleton clusters. Communities in each of the consensus matrix (D) are identified by applying genLouvain community detection 100 times on co-association matrix and by considering the community with maximum modularity score (Q).

Step-4: Ranking Genes for Inter-layer Graphs (I_2):

We use consensus communities found using the genLouvain method (Table TC5.3), and rank the genes based on their key roles in each intra-layer graph. The consensus communities are treated in two categories: 'Highly connected components' (HCCs) and 'Non-conformers components' (NCCs). If the total number of nodes in a community is greater than the average of nodes of all non-singleton communities, we consider that community as an HCC; else, the community falls under the NCC category. The HCCs found in the breast cancer data expression traits layer and methylation feature layer is 79 (out of 777) and 42 (out of 130). Each *HCC* of each layer is collected from the respective correlation layers and ranked the genes using their node degree, node betweenness centrality, and eccentricity measures. The highest betweenness centrality score implies that the node is part of the shortest path for most nodes in the network. The node's whose eccentricity is equal to the radius of the network is considered as a central node of the network. We thus identify three sets of nodes in each community in each intra-layer graph, (1) all central nodes in the network, (2) top 10% of highly ranked genes based on their node degree, and (3) top 10% of those based on betweenness centrality. The union-set of these three sets gives us the significant genes in the layer. Each layer's NCCs are clubbed together and treated as an HCC, and then ranked the genes. The total ranked genes of expression traits and methylation features for breast cancer are 3895 and 1882, respectively, ranked genes of lung cancer are 3403 and 3061, respectively (Table TC5.2, Column *Step-4*).

Step-5: Construction of Inter-layer Graph in `HCNM`:

The inter-layer graph is computed using Spearman's rank correlation [179] between the selected genes from **Layer-1** and **Layer-2**. Unlike Pearson's correlation measure, this method is not a linear correlation measure, also it estimates the coefficient value between the variables without any frequency distribution requirement [180] and is less sensitive to outliers. We have used Spearman's procedure, as we expect a monotonic relationship between expression traits and methylation features. The Spearman correlation between the variables is:

$$\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)} \qquad \text{(Eqn 5.10)}$$

Where d_i is the difference measure of the ranks of variables of n samples data.

The submatrix/bipartite-graph of breast cancer is of [3895 × 1882], and lung cancer is [3403 × 3061] (Table TC5.2, Column *Step-4*). We then further sparsify this graph using a p-value and threshold from PA, as done for intra-layer graphs. We find $\tau = 0.2$, $\tau = 0.3$ in our case study for breast and lung cancer inter-layer graphs, respectively. This completes the construction of our proposed `HCNM`.

Step-6: representative Integrative Subspace:

To further reduce the number of significant genes to a representative set, we rank the gene-pairs in the inter-layer graph based on their edge-betweenness centrality measure.

The top 10% of these pairs that had also participated in most of the shortest paths of the network are finally selected. We find the integrative subspace of these selected genes that are feature-rich and significant. Our representative integrative subspace now has 531 expression traits and 339 DNA methylation data of genes for 486 samples for breast cancer and 904 expression traits, and 785 DNA methylation data of genes for 200 samples for lung cancer data (Table TC5.2, Column *Step-6*).

5.3 Biological Significance of Subspace in Genes

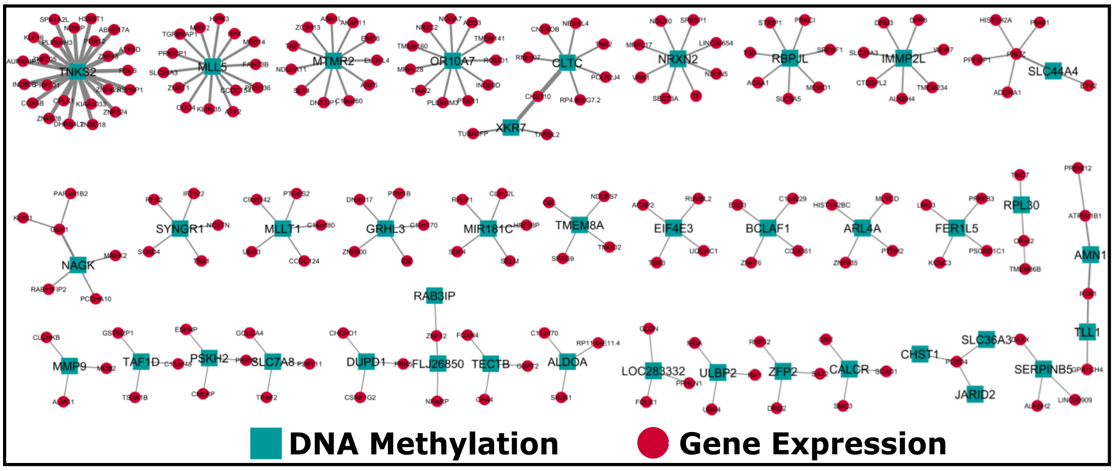


Figure FC5.2: The identified significant ranked genes of breast cancer. The connected components of the graph, with 3+ nodes, and of genes retained after edge (gene-pair) ranking, with edges between expression traits (red circular glyphs) and DNA methylation (turquoise square glyphs) data of genes, and the edge width indicates the edge betweenness score. Note - the gene symbols can be read at 200% zoom-in of the document.

Our model demonstrated new and interesting characteristics of the identified genes in subspace. We observe that the bipartite inter-layer graph creates several connected components, forming like star-graphs of expression traits genes around methylation features. Such graphs indicate that there is an association between several expression traits to single methylation feature, as observed in star graphs with 3+ nodes in our case study in Figure FC5.2. Interestingly we can also observe that all three-node connected components are with methylation features as a central node, which can be attributed

Table TC5.4: The top 10 enriched GO terms of the ranked genes.

GO	**Term**	N_G	**PValue**
(i) Breast Cancer			
UP_KEYWORDS	Phosphoprotein	**382**	5.33E-08
UP_KEYWORDS	Alternative splicing	**451**	4.91E-05
GOTERM_MF_DIRECT	GO:0005515∼protein binding	**394**	1.13E-04
INTERPRO	IPR013164:Cadherin, N-terminal	11	1.93E-04
UP_KEYWORDS	Transit peptide	38	2.78E-04
GOTERM_CC_DIRECT	GO:0005737∼cytoplasm	**245**	6.07E-04
UP_SEQ_FEATURE	splice variant	**339**	7.29E-04
GOTERM_CC_DIRECT	GO:0005829∼cytosol	**164**	8.24E-04
UP_KEYWORDS	Cytoplasm	**218**	0.001086
UP_SEQ_FEATURE	domain:Cadherin 6	11	0.001183
(ii) Lung Cancer			
UP_KEYWORDS	Phosphoprotein	**710**	1.92E-10
UP_KEYWORDS	Alternative splicing	**869**	6.40E-09
UP_KEYWORDS	Metal-binding	316	0.000142
UP_SEQ_FEATURE	splice variant	**635**	0.000186
UP_KEYWORDS	Zinc	211	0.000441
BIOCARTA	Arrestin-dependent Recruitment of Src Kinases in GPCR Signaling	8	0.000449
GOTERM_BP_DIRECT	GO:0006357∼regulation of transcription from RNA polymerase II promoter	54	0.000572
UP_KEYWORDS	Zinc-finger	164	0.000721
GOTERM_MF_DIRECT	GO:0008093∼cytoskeletal adaptor activity	7	0.000794
GOTERM_MF_DIRECT	GO:0022857∼transmembrane transporter activity	12	0.000944

to methylation features that are regulating the gene expressions. We observe similar patterns with the lung cancer ranked genes network.

Feeding these ranked genes (Table TC5.2, Column *Step-6*), i.e., 870 genes (531 expression traits + 339 methylation features) of breast cancer and 1689 genes (904 expression traits + 785 methylation features) of lung cancer of our integrative subspace into ‘Database for Annotation, Visualization and Integrated Discovery (DAVID)’ tool [181,182], we get the enriched gene ontology (GO) terms, with N_G genes belonging to each term, as listed in Table TC5.4. The subspace in genes in both the cancer profiles had resulted in enriched gene ontology terms such as alternative splicing, splice vari-

Table TC5.5: A few top genes associated with breast and lung cancer, ranked by their Gene-Disease Association scores are obtained from the DisGeNET database.

Disease	**Disease ID**	**Gene**	**GDA Score**
Breast Carcinoma	C0678222	STAT3	0.4
		MAPT	0.2
		ATF2	0.08
		TRAF2	0.07
		NUMA1	0.03
Malignant Neoplasm of Breast	C0006142	STAT3	0.4
		ATF2	0.37
		PPHLN1	0.3
		UBR4	0.3
		RFX2	0.3
Triple Negative Breast Neoplasms	C3539878	STAT3	0.1
		SPAG9	0.02
		DAXX	0.01
		CRTC1	0.01
Squamous Cell Carcinoma of Lung	C0149782	TTN	0.31
		ATR	0.3
		GRAMD4	0.01
		CLDN7	0.01
		CACNA2D3	0.01
		SRPK1	0.01
		TTN	0.01
		KMT2B	0.01
		RGS3	0.01
Lung Diseases	C0024115	ENG	0.31
		MYLK	0.03
		DHPS	0.01
		RASSF1	0.01
		ACD	0.01
		LAIR1	0.01
		IFNAR2	0.01
		HDAC2	0.01
		BEST1	0.01
		APC	0.01

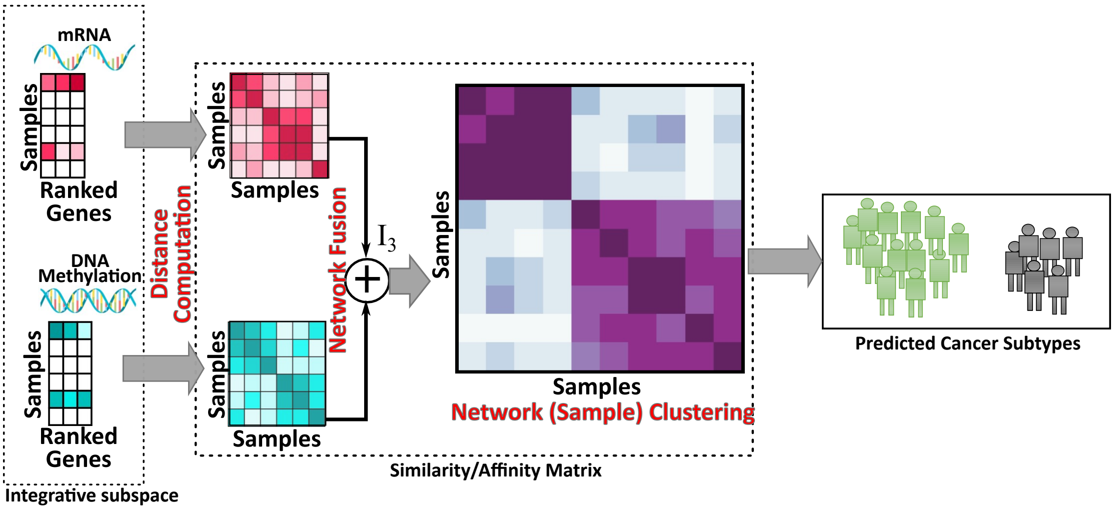
ant, protein binding, phosphoprotein, regulation of transcription, metal-binding, etc., (Table TC5.4). These top 10 terms are sorted based on their false discovery p-value in Table TC5.4. We have verified the gene-disease association (GDA) score of the shared genes found from the GO terms with $N_G > 25\%$ of total genes (highlighted in column N_G) using the DisGeNET database (https://www.disgenet.org/). A total of 38 and 346 genes have been found in common across top enriched GO terms of breast and lung cancer, respectively. 35 out of these 38 genes of breast cancer have positive GDA scores when studied using the DisGeNET database (https://www.disgenet.org/), implying the presence of evidence in the literature indicating the association of the gene with the disease. Few top-ranked genes associated with different breast and lung cancer subtypes are given in Table TC5.5. Our HCNM successfully identifies the significant feature-rich subspace in genes of both the cancer profiles, that are biologically significant.

5.4 Conclusions

Our proposed HCNM model successfully identified the subspace in genes in both breast and lung cancer datasets. Heterogeneous correlation network analysis adds to our model as both independent and cross-correlations among the genes are effectively utilized. Our multilevel integrative model greatly decreased the dimensions of data by considering both consensus communities and network topology. Overall, we can conclude that our HCNM successfully identifies the significant feature-rich representative subspace in genes, which are optimal for finding the disease subtypes. In our next chapter, we use the gene-subspace, for an application of cancer subtype prediction. We use multi-omics data of each cancer profile to identify the subtypes in samples/patients using multi-omics integrative procedures such as SNF, ANF, and iCluster.

CHAPTER 6

SUBTYPES IN CANCER



Prediction of cancer subtypes is vital for early prognosis of the disease, for personalized treatment, and improved likelihood of survival of the subject. We study the subtypes in breast cancer data, one of the well-studied cancer phenotypes, and subtypes in lung cancer data, one of the least studied phenotypes. In this Chapter, the subtypes of each phenotype are found using subspace in genes and by applying multi-omics integrative studies for subtype prediction. We compare the subtypes found using subspace in genes with subtypes of entire data. We also compare the subtypes using the available popularly known subtypes annotation in the literature, i.e., *PAM50* labels for breast cancer subtypes [183, 184], and *expression subtypes* labels for lung cancer [185].

Here, we propose a two-step algorithm to predict subtypes for a cancer phenotype. (1) Integration of sample similarity networks across different omic features by network fusion I_3 (Figure FC1.2), and (2) clustering of the fused similarity network to find subtypes. Steps (1)-(2), including I_3 can use several network fusion methods [161, 164, 186, 187].

The details of the representative integrative subspace of multi-omics data are provided in Section 6.1, integrated multi-omics algorithms were used to find disease subtypes are described in Section 6.2, and the comparison of results across different methods are provided in Section 6.3.

6.1 Dataset

Table TC6.1: Data dimensions before and after finding representative *integrative subspace* of multi-omics in breast and lung cancer profiles. The annotated subtype categories of each cancer profile are provided in the column 'Subtypes'.

	After Preprocessing		**Subspace Data**		**Subtypes**
	mRNA	*Methy.*	*mRNA*	*Methy.*	
Breast Cancer Samples = **486**	16087	10040	**531**	**339**	*PAM50 Subtypes* [183] 1. Luminal A 2. Luminal B 3. HER2 Positive 4. Triple-Negative 5. Normal
Lung Carcinoma Samples = **200**	16877	9405	**904**	**785**	*Molecular Subtypes* [184] 1. Basal 2. Classical 3. Primitive 4. Secretory

To find disease-specific patient subtypes and to validate the subspace in genes, we have used a) the entire data, i.e., preprocessed data of methylation features and expression traits (Section 5.1), and b) the representative subspace data produced by multilevel integrative procedures I_1 and I_2 (Chapter 5). We compare our results of the entire data with subspace data. The dataset dimensions are given in Table TC6.1. The subspace in breast cancer data is 3.3% of the total (16087) expression traits and 3.38% of the total (10040) methylation features. Similarly, the subspace in lung cancer data is 5.36% of the total (16877) expression traits and 8.35% of the total (9405) methylation features (Table TC6.1). Overall, in both the phenotypes, the representative subspace is less than 10% of total genes. The benchmark annotated subtypes of breast and lung cancer phenotypes, i.e., PAM50 subtypes and molecular expression subtypes, respectively, are downloaded using R, TCGAbiolinks library [163]. The classified breast cancer subtypes of PAM50 are of five categories, namely, luminal A, luminal B, HER2 positive, triple-negative or basal-like type, and normal categories, and the molecular expression subtypes of lung cancer are categorized as basal, classical, primitive, and secretory (Table TC6.1, column-‘Subtypes’).

6.2 Subtype Prediction using Representative Integrative Subspace

We use network-based (NB) multi-omics fusion integrative methods for patient/samples subtype prediction. We first identify the similarity or affinity networks of the samples for each omic feature in the integrative subspace, and then fuse them as an integrative step (I_3 in Figure FC1.2). We then find clusters of samples representative of cancer subtypes.

Step-1: Network Fusion (I_3*)*:

Most multi-omics integrative algorithms, such as similarity network fusion (SNF) [164], and affinity network fusion (ANF) [186] integrate data of multi-omic after computing similarity or affinity matrices internally. ANF is an improvised integrative procedure on SNF; both the methods first compute the distance between patients and using the distance matrix, affinity measurement of each genomic feature is computed separately, then using network fusion procedure, a final multi-omics integrated network is generated. We use SNF and ANF in our work, by tuning the hyperparameters K (number of neighbours), σ (variance for affinity measurement), α (measure for local diameter) and β (measure for pair-wise distance) using the correlation measure (*ref:* Equation 7 [42]). The number of clusters is estimated using eigen gap and rotation cost methods.

We also use iCluster [188], a multi-omics integrative method where similarities between the samples and clusters are computed simultaneously by minimizing the intra-cluster variance. We use the 'tune.iClusterplus' method to find the optimal number of clusters and Lasso penalties. Compared to SNF and ANF, the computation time taken for the iCluster procedure is higher because the implementation of iCluster directly gives the clusters, thus combining step-1 and step-2.

We have implemented these methods using R packages, namely, SNFtool [164], ANF [189], and iClusterPlus [190]. The detailed implementation specifics of these multi-omics integrative algorithms, i.e., SNF, ANF, and iCluster, are provided in *Appendix B*. The implementation of SNF library as a tool using *R-Shiny* is available at GitHub repository, `https://github.com/vrrani/SimilarityNetworkFusion`.

Step-2: Sample Clustering for Subtype Prediction:
In order to find clusters in the fused similarity or affinity networks, we extract clusters of samples using spectral clustering, using the estimated number of clusters, done in Step-1. We then compare these clusters or subtypes with the popularly known subtype annotation data of TCGA, namely `PAM50` [183,184] for breast cancer, and `expression subtypes` [185] for lung cancer.

6.3 Comparative Analysis of Subtypes

Table TC6.2 is the list of the estimated number of subtypes/clusters of both phenotypes. Amongst the integrative multi-omics procedures considered here, the computation time for iCluster is significantly more than the other methods; the same is observed in (*ref:* Figure 3(c), [164]). Hence, iCluster is not implemented on full gene space.

Table TC6.2: Count of estimated clusters/subtypes for each of the integrative multi-omics algorithms on representative subspace $(*\_HCNM)$ data and entire data $(*\_Full)$ of breast and lung cancer phenotypes.

Disease Phenotypes →	**Breast**	**Lung**
Methods ↓		
SNF_HCNM	3	3
ANF_HCNM	4	3
iCluster_HCNM	4	3
SNF_Full	3	3
ANF_Full	4	3
Annotation Subtypes	5 [183, 184]	4 [185]

Using subspace in genes data and multi-omics integrative procedures, we have obtained the subtypes and compared them with subtypes annotation *PAM50* of breast cancer and *ExpressionSubtypes* of lung cancer data. The plotted Kaplan-Meier survival curves of breast cancer (Figure FC6.1) displayed a clear separation between good survival and poor survival subtypes based on their survival probabilities in fusion-based models SNF and ANF with subspace data. This also echoes our assertion that reduced dimension (representative subspace) data is more significant than the entire (full) data of each genomic feature, as subspace data is less prone to noise, bias, and outliers. The SNF method on representative subspace data displayed better subtypes when tested at median survival-probabilities (50%) with the least log-rank p-value, i.e., 0.094. A similar pattern is observed with lung cancer subtypes. The subtypes found using subspace in genes are comparable to the subtypes found with complete multi-omics feature space. The benefit of using the reduced dimensionality data is that the subspace is less prone to noise, bias, and outliers. But, though the dataset contains 200 subjects for lung cancer and 486 subjects for breast cancer, the available annotation information for subtypes is only for 54 and 480 subjects for lung and breast cancer data, respectively. Hence, for subtypes comparison, we have considered only 54 subjects for lung cancer, and 480 subjects for breast cancer.

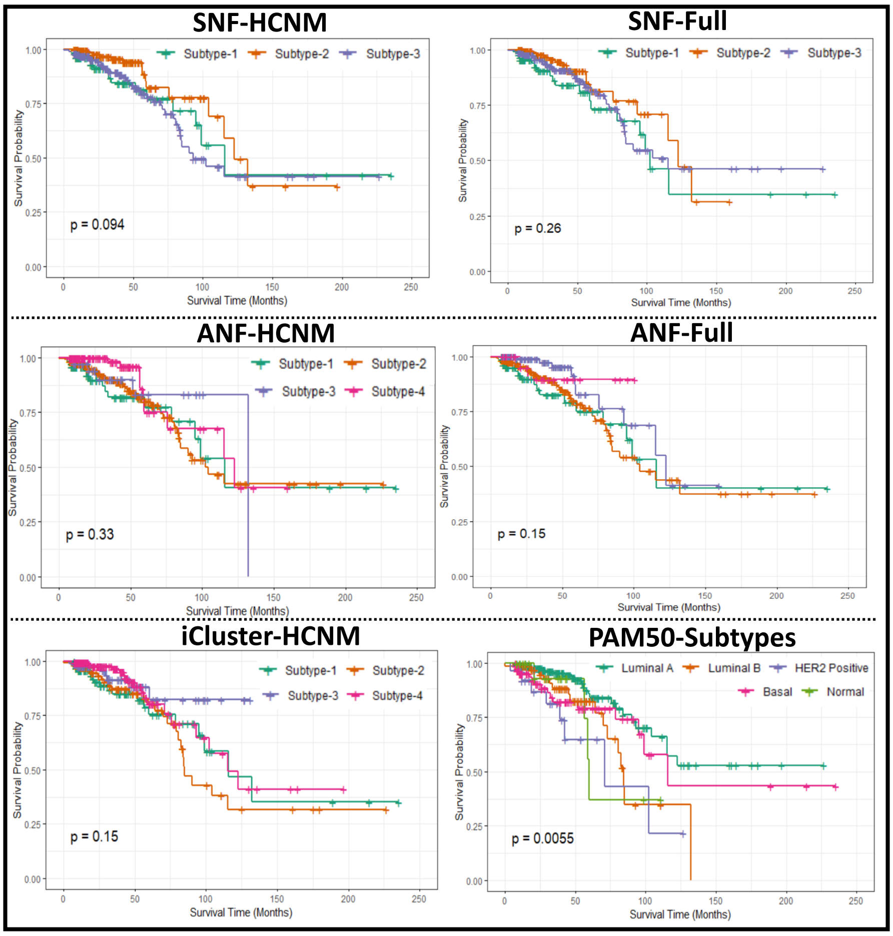


Figure FC6.1: The good and poor survival times of breast cancer data for subtypes were predicted using different methods. Subtypes are significant at median survival-probability in all methods. We see clear survival probability separation for subtypes identified using SNF with subspace in genes and benchmark study using annotated subtypes, *i.e.*, `PAM50`, than using ANF and iCluster.

Comparing the subtypes of breast cancer using the Sankey plot depicts more agreement between SNF and ANF subtypes (Figure FC6.2). The subtypes in SNF_HCNM (representative subspace) model divides into two subtypes in ANF_Full model (second and third vertical bars in Figure FC6.2). We also observe similar behavior when compared the patients subtypes with annotation subtypes, using Jaccard similarity (S_j) and *NMI*. The S_j and *NMI* are slightly higher with subspace data and fusion-based methods, i.e., with SNF or ANF compared to full/entire data, assuring the subspace in genes identified using `HCNM` is significant (Table TC6.3). Overall, the network fusion based

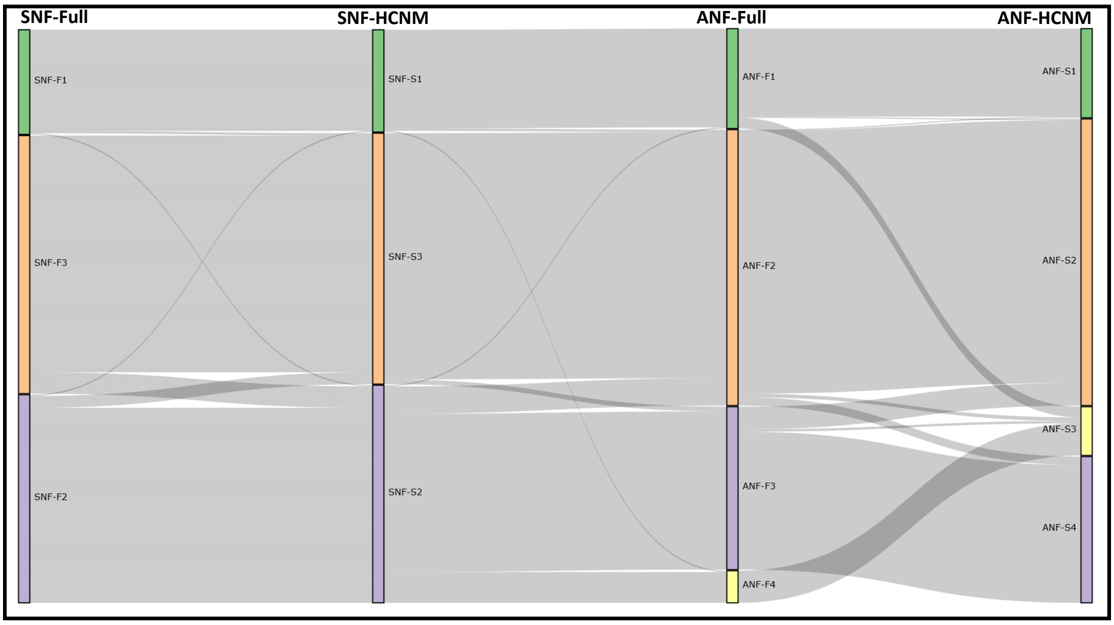


Figure FC6.2: Sankey plot of the patient subtypes from using our algorithm using network fusion methods (SNF, ANF) with data from the complete (Full) gene space and our representative integrative subspace (HCNM), shows that subtypes found using SNF-HCNM data agree with SNF/ANF-Full data, more than ANF-HCNM.

methods are more favorable over iCluster, when used with our representative subspace, owing to the computational time, and the results in Table TC6.3 and Figure FC6.1.

Table TC6.3: Quantitative measures to compare the subtypes.

Scores	*SNF* (HCNM)	*SNF* (Full)	*iCluster* (HCNM)	*ANF* (HCNM)	*ANF* (Full)
(i) Breast Cancer					
NMI	**0.43**	0.42	0.33	0.39	**0.43**
S_j	**0.46**	0.44	0.42	0.51	**0.51**
(ii) Lung Cancer					
NMI	0.37	0.39	0.32	**0.45**	0.42
S_j	0.64	0.66	0.64	**0.71**	0.71

6.4 Conclusions

We have used the representative integrative subspace data identified using our heterogeneous correlation network model HCNM, to find the patient subtypes in breast and lung cancer phenotypes. The fusion-based integrative multi-omics methods are most suitable for finding the patient subtypes. Using the quantitative and qualitative assessments, we infer that the subtypes found using subspace data are more in agreement with the benchmark annotated subtype labels in literature. The HCNM model is efficient in finding subspace in genes that are feature-rich and significant for each phenotype, as subtypes found with subspace in genes displayed better survival prediction plots, log-rank p-value, *NMI*, and S_j when compared with full/complete data. For evaluating the overall performance of the method, we need to perform an ablation study, which is in the future scope of this work. We intend to integrate Bayesian methods more extensively into multi-level integration algorithms to improve combining network-based and Bayesian methods as future work. The extensibility and scalability of our algorithm to other omic features require further in-depth study.

Part III

Extensions to Network Analysis

CHAPTER 7

MATRIX VISUALIZATION APPLICATIONS

In this chapter, we describe the extended applications of biomedical data using matrix visualizations. The first study focuses on visualizing FC matrix derived from rs-fMRI data. The FC is a complete (unthresholded) correlation matrix that depicts the inter-correlated clusters when an appropriate seriation technique is employed. The second study describes a novel hybrid graph layout, named **RadTrix**, that is proposed to visualize an unbalanced bi-partite graph. RadTrix is a composite of *matrix* and *radial/circular* visualization. We have used a disease-gene association network, the 'diseasome', as a case study for the RadTrix layout.

7.1 Matrix Seriation for Visualizing Changes in FCN

Brain functional connectivity network is a fully connected undirected graph that can also be represented as an adjacency matrix. Visualizing the adjacency matrix using seriation, *i.e.*, reordering, rows and columns helps reveal the block-like structures along the diagonal, which are patterns for clusters. Matrix seriation or permutation is an approach that naturally reveals the cluster information in the network, which doesn't necessitate any prerequisites such as sparsifying the network or providing an input parameter to find the clusters. Bertin defined seriation as a procedure of 'simplifying without destroying' [191]. The correct choice of reordering an adjacency matrix naturally reveals the semantically relevant clusters in the matrix. There are several methods to implement seriation [158], such as, principal component analysis (PCA), visual analysis for cluster tendency Assessment (VAT) [192], Random and Rank-two ellipse (R2E) seriation [193], etc. These seriation procedures have been used to examine the cortical structure of mammalian brains [194] and in genomic applications [195].

7.1.1 A Case Study of an OCD Patient fMRI Data: Changes Due to Treatment

We have used a matrix seriation procedure to understand the patterns of brain regions of an obsessive-compulsive disorder (OCD) subject data [196]. The subject is a 21 years-old male student with declined academic progress and severe obsession, compulsion, and mentally disturbed symptoms, including suicidal thoughts, etc., along with motor coordination issues such as problems regarding vigilance, execution, verbal, etc. The MRI images exhibited a lesion at the posterior cerebellum at the right side of the brain (Figure FC7.1) (i), which can be due to the obstruction of the blood supply in the posterior inferior cerebellar artery region. The pre-supplementary motor area (pre-SMA) region connectivity with the lesion region is depicted in Figure FC7.1 (ii). The cross-hairs on Figure FC7.1 (i) depict the lesion region, and the connectivity of the lesion with pre-SMA is shown in Figure FC7.1 (ii). Hence the OCD symptoms are presumably due to secondary to posterior cerebellar infarct in right crus II. The fMRI data of post and pre-treatment of this subject is used to demonstrate the enhanced network modularity of the cerebellar network, with improved symptoms after treatment.

The rs-fMRI data of this subject is studied before and after ‘repetitive transcranial magnetic stimulation’ (rTMS) treatment. rs-fMRI scans of 250 volumes are acquired using a 3-Tesla scanner with a 20-channel coil at National Institute of Mental Health & Neurosciences (NIMHANS)[1], with the consent of the subject. The raw fMRI files

[1] `https://nimhans.ac.in/about-us/`

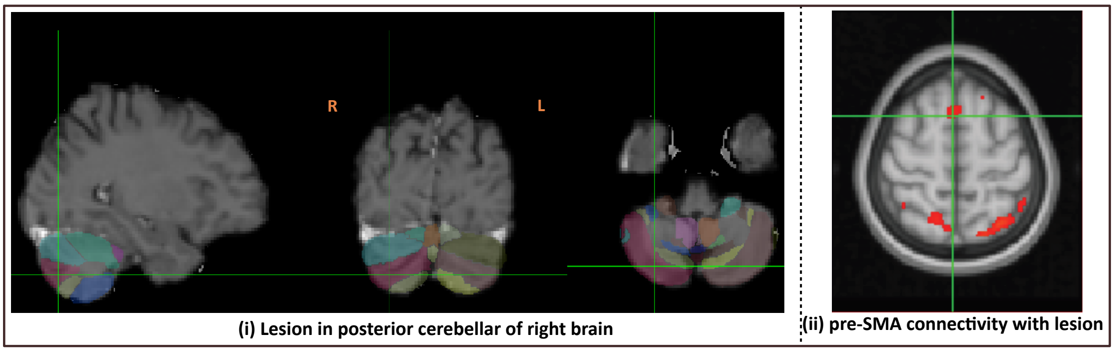

R
L
(i) Lesion in posterior cerebellar of right brain
(ii) pre-SMA connectivity with lesion

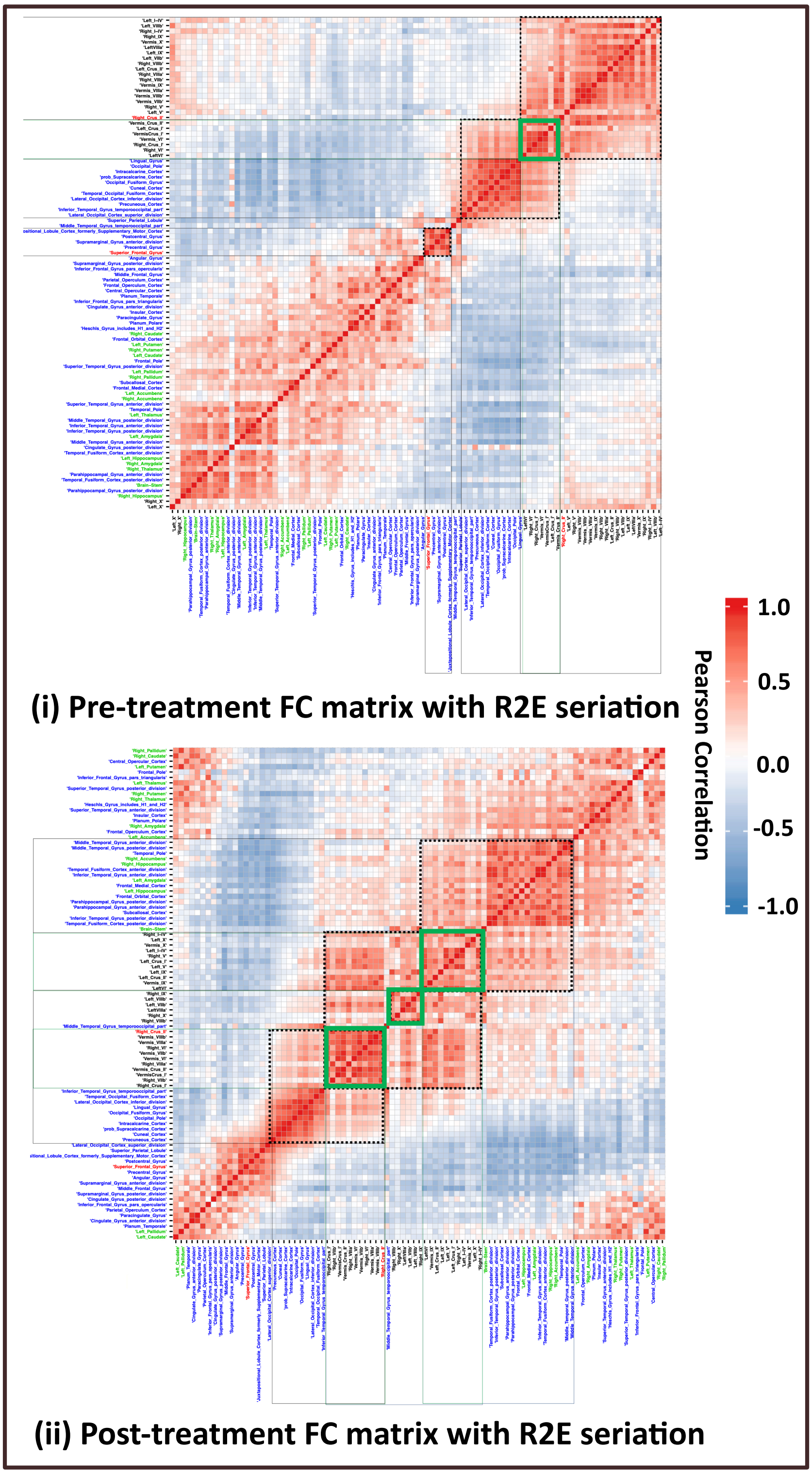
Figure FC7.1: Brain fMRI scans that demonstrate (i) lesion at posterior cerebellar right crus II (cross-hair marks) and (ii) pre-supplementary motor area (pre-SMA) region connectivity with the lesion regions.

of pre- and post-rTMS treatment are available at Harvard Dataverse[2]. The scanned images are processed using the FMRIB software library of version-5.0.10 [197]. The pre- and post-treatment fMRI scans are subjected to Harvard-Oxford (HOA) atlas [198] and MNI-FLIRT atlas [199]. The total analyzed ROIs are 48 from the cortical area, 15 from the subcortical area, and 28 from cerebellar regions of the brain. The BOLD time-series signals of these selected 91 ROIs/nodes are aggregated, and pairwise Pearson's correlations are measured across the 91 nodes (48+28+15). Hence, after pre-processing the data, two FC matrices of $[91 \times 91]$ are generated for each pre- and post-treatment scan. The correlation strengths of the two $[91 \times 91]$ FC matrices were studied using 'Rank-two ellipse' (R2E) seriation procedure that carefully sorts items into an ordered series to identify recognizable patterns.

7.1.2 Rank-two Ellipse (R2E) Seriation

Seriation techniques reorder nodes to improve the clarity of blocks along the diagonal by moving the higher correlation valued cells closer to the diagonal, i.e., Robinson matrix order [200]. It is the process of identifying new ordering or permutation of rows and columns to reveal the patterns and finer structure such as: subnetwork, cliques,

[2] https://doi.org/10.7910/DVN/X12BZD



(i) Pre-treatment FC matrix with R2E seriation

(ii) Post-treatment FC matrix with R2E seriation

Figure FC7.2: Matrix visualization after applying rank to ellipse (R2E) seriation (i) before rTMS treatment and (ii) after rTMS treatment. The cerebellar network and other connected networks are represented with black-dotted bounding boxes, and inter-network overlaps are depicted with green colored bounding boxes. A single overlapped region before treatment (green box) has switched across three distinct overlapped regions after treatment, signifying improved modular network structure. Cerebellar, cortical, and subcortical nodes are given in black, blue, and green colors, respectively. The lesion node, i.e., 'right_crus_II' and the neuro-stimulation region, 'superior_frontal_gyrus' are represented in red color.

clusters, central actor/hub nodes, etc. Thus, along the diagonal, the matrix visualization shows the clusters in the network. For reordering, we use a seriation method on the matrix, which is a two-way one-mode, i.e., the matrix uses the same permutation order of the nodes along its rows and columns. Thus, matrix seriation is a permutation method.

Here, we have implemented the R2E seriation technique on the two $[91 \times 91]$ matrices for node clustering [193] (Figure FC7.2). R2E seriation involves iteratively finding the correlation of correlation matrices of p nodes by considering columns in the matrix as p dimensional points, and this iterative process has been found to converge onto an ellipse in two-dimensional space. R2E seriation at the n^{th} iteration, when converged, all elements of the matrix will be ± 1, and the observation of points being laid out on the ellipse, at which the matrix reduces its rank from p to 2. Hence, the clusters revealed in the seriation are found to be laid out as the clusters on this ellipse. The unique positions of p points on the ellipse are used to find the reordering of rows and columns, representing the correlation matrix in the Robinson matrix format. The Robinson form, for a symmetric matrix, implies $a_{ij} \leq a_{ik}$ for $i < j < k$ (upper-triangle matrix). The advantage of the R2E procedure is that its low execution time, compared to other seriation methods, since convergence is guaranteed within fewer iterations [201], with smoother transitional patterns that are biologically meaningful [202].

On visualizing the changes on the seriation maps before and after the treatment using FC matrices, we have observed (a) extended connectivity of the cerebellar network (black nodes) – the larger cerebellar cluster/block had an increased overlap with both anterior and posterior brain networks as observed along the diagonal as shown in Figure FC7.2 (ii), and (b) formation of better-defined sub-clusters within the larger cerebellar cluster, indicating improved within-network modularity of distinct functional cerebellar networks. That is to say, the connectivity of the cerebellum before treatment that was predominantly with the occipital brain regions changed post-treatment to reveal more significant coactivation with the parietal, temporal, and frontal regions. Besides, we could observe more distinct modularity within the cerebellar nodes post-treatment, with the vestibular network (lobules IX and X) separating from the cognitive-limbic network (crus I/II and vermis), while remaining within the larger cerebellar cluster. This indicates that stimulating the pre-SMA region could have improved the within- and between network connectivity of the posterior cerebellar brain regions and thus driven the change in the clinical profile of the OCD subject.

7.2 A Hybrid Graph Layout for Unbalanced Bipartite Graph

A bipartite graph, also known as bigraph or 2-mode network, is a graph with two disjoint sets of nodes laid in parallel lines, with the inter-set links connecting these disjoint sets. The bipartite graphs are viewed as either vertical or horizontal, two-layer graph layout and are used to study networks of biomedical, biomolecular, epidemiological, ecological, protein complexes, etc [203]. The bipartite graphs are generally visualized using node-link layouts, but these graphs are highly prone to visual clutter, especially when the sets of nodes are unbalanced, i.e., the cardinality of nodes in one set of the bipartite graph is much greater than the other set (Figure FC7.3). Hence, we propose a novel hybrid visualization layout, **RadTrix**, for a bipartite graph with a relatively large skew in the vertex set sizes in the two sets.

7.2.1 RadTrix Layout

RadTrix is a *composite visualization*, specifically of the type of *nested views* [203]. It is a composite of two visualizations in a single view, i.e., matrix visualization for a smaller set and circular/radial graph layout for the larger set of nodes in a bipartite graph. The design pattern used here is the choice of visualization techniques for the two sets. The use of radial layout allows for a layout of the larger set where the screen space is maximally utilized. The use of the matrix layout uses its quadratic spatial complexity

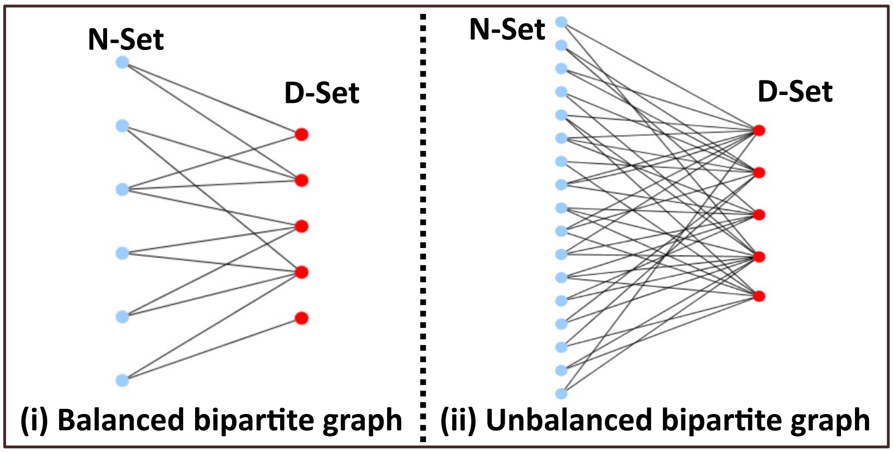


Figure FC7.3: A bipartite graph of a vertical two-layer layout with the nodes of (i) balanced cardinality (ii) unbalanced cardinality.

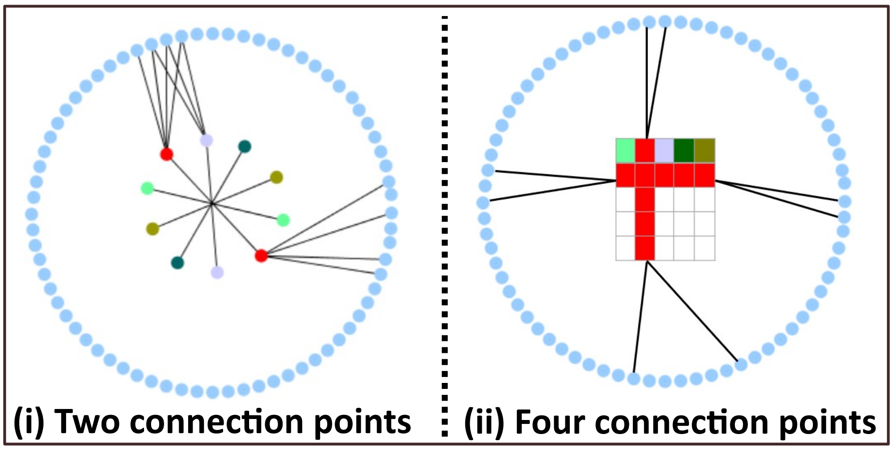


Figure FC7.4: A toy example for comparison of choice of layout for D nodes to reduce the visual clutter. A graph layout with (i) two connection points, (ii) four connection/landing points.

to facilitate four landing spots for nodes, i.e., endpoints for edges incident on nodes, in the smaller set, thus reducing the clutter due to crowding of edges (Figure FC7.4). Compact and intuitive visual representation is achieved in this design, as is expected from nested views [203]. At the same time, the disadvantage of limited space for matrix visualization remains, thus, limiting the size of the smaller set.

The radius r of the radial layout is based on maximizing the use of screen space for visualization. The N nodes of the larger-set are uniformly placed on the circumference of the circle, with an angle difference of $\theta = 2\pi/N$ between them. The location for each of the Nodes in the N-set are placed using the polar coordinates, relative to the center of the circle, $(r\cos(k\theta), r\sin(k\theta))$ for $k \in \mathbb{Z}$ and $0 \leq k < N$ to place the N nodes in the circumference. We recommend $r \geq 4 \times N$. Finally, for rendering links between the node on the circumference and a node in the matrix, our algorithm compares the distance of the location (obtained from the polar coordinates) of the circumference node with each of the four landing spots of the matrix node, and uses the landing spot giving the minimum distance to draw the link between the nodes. We have implemented the RadTrix algorithm using D3.js library. The tool is available at the GitHub repository `https://github.com/vrrani/RadTrix`.

We describe the RadTrix layout construction using the following steps:

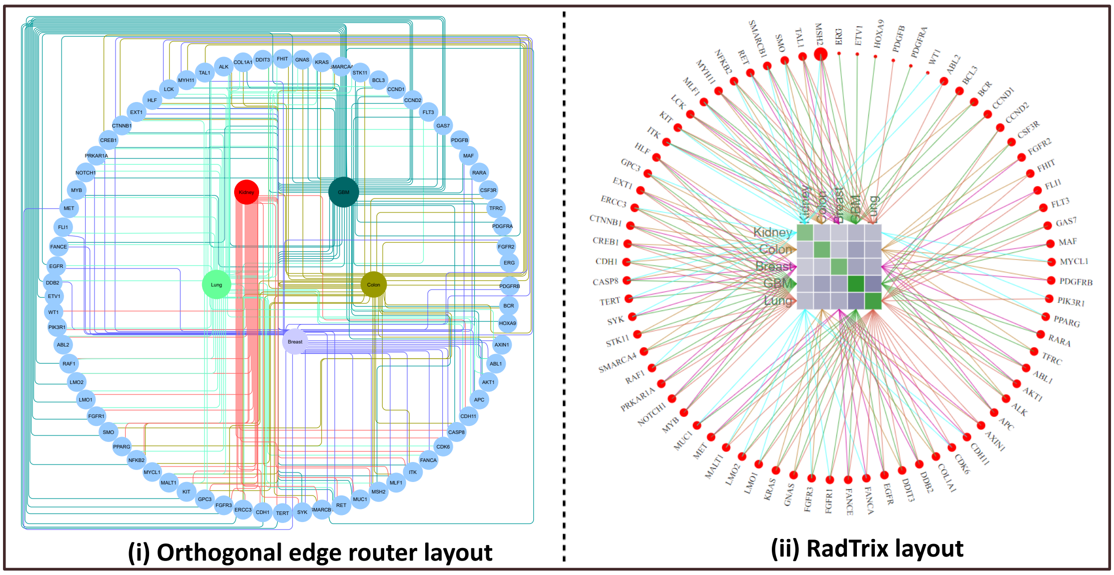
1. Arrange the larger set of nodes of a bipartite graph uniformly on the circumference of the radial layout.
2. Draw the nodes on the circumference as per the size corresponding to its property, for example, node degree.
3. Arrange the smaller set nodes as a matrix, with the same ordering of nodes used in rows and columns.
4. Color the matrix cells based on a specific criterion of the relationship between the nodes.
5. Merge the radial and matrix layouts by placing the center of the latter with that of the former, and scale the matrix to fit inside the circle of the former.
6. Render the links from nodes on the radial layout to the nearest landing spot of the nodes in the matrix layout. (The four landing spots are the left and right side of the row representing the node in the matrix layout; and the top and bottom side of the column representing the same.)

7. Re-order/seriate the nodes in both the matrix and the radial layout to visualize the patterns.

The beneficial feature of RadTrix in reducing visual clutter lies in the design pattern of exploiting the quadratic spatial complexity of the matrix layout to provide with four landing spots. In comparison, a radial layout gives two landing spots per node (Figure FC7.4 (i)), and a traditional node-link diagram gives only one. Reducing visual clutter implies the reduction of edge crossings which improves the readability of the unbalanced bipartite graph.

7.2.2 A Case-study using RadTrix

As a case-study, a diseasome network, i.e., disease-gene association network is used where diseases and genes are the two disjoint sets of the bipartite graph. Disease-gene association networks are well studied as a bipartite graph [204–206]. One of the challenges in visualizing the bipartite graph in this specific scenario is that generally, for D diseases, we have N genes in the diseasome, where $D \ll N$. The difference in set cardinalities in such an *unbalanced bipartite graph* leads to visual clutter in the case of two-layer graph layout representation [207] (Figure FC7.3 (ii)). Hence, we use the RadTrix layout to visualize and understand disease-gene associations. The diseasome network is generated using an integrated multi-omics data in [208]. We have constructed the diseasome using the set of 35, 36, 58, 29, and 51 genes for `Breast`, `Colon`, `GBM`, `Kidney`, and `Lung` cancer profiles, respectively. These genes are common in both mRNA and DNA methylation genomic networks of integrated analysis in [208]. This gives us 209 genes in all; however, since there are genes shared across different disease phenotypes, we have removed duplicates and reduced the set of unique genes in the diseasome to 73. Hence the final diseasome network is with 5 diseases (D) and 73 genes (N), 209 links between D and N (edges), with set cardinalities as, $D \ll N$. We have a 5×5 matrix layout of the five disease phenotype nodes, and the 73 genes uniformly placed on the radial layout.



(i) Orthogonal edge router layout

(ii) RadTrix layout

Figure FC7.5: The diseasome of our case study, generated using (i) Orthogonal Edge Router layout of the yFiles, in Cytoscape [5]. (ii) Our proposed RadTrix layout.

To explore *which* genes are significant between the diseasome phenotypes, we use the RadTrix layout of the diseasome. Similar to RadTrix, the Orthogonal edge router layout uses radial layout for the genes and represents the circle interior to place the disease nodes. However, this layout is not optimized to reduce edge crossings, leading to visual clutter (Figure FC7.5 (i)). Hence, we have performed gene-centric analysis using RadTrix, as shown in Figure FC7.6.

The matrix layout of D diseases follows the *two-way one-mode* format [158], i.e., representing a $D \times D$ matrix $\mathbb{D}$, where rows and columns refer to the same entities in the same permutation order. We use the matrix cells to visually encode a unary operator in the diagonal elements and a binary operator in the non-diagonal elements of the matrix. For example., we can represent diagonal element $\mathbb{D}_{ii}$ as the number of genes associated with disease D_i, and non-diagonal element $\mathbb{D}_{ij}$ as the number of common genes asso-

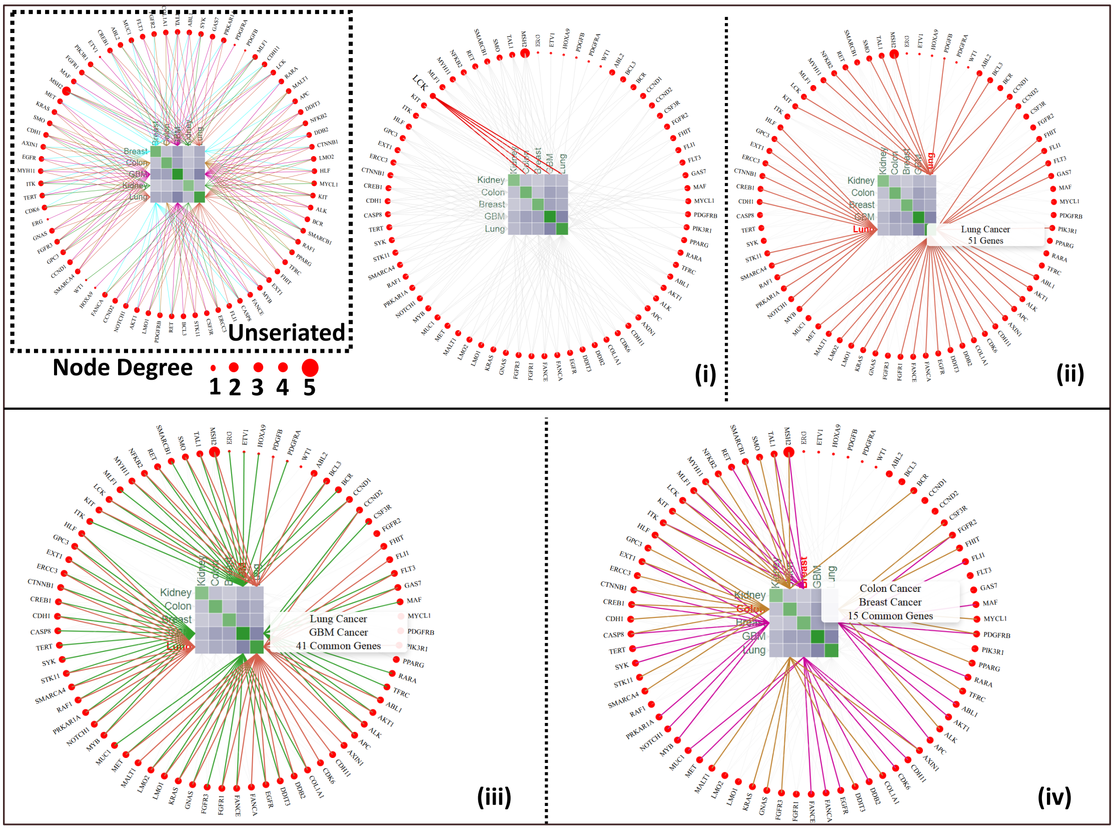


Figure FC7.6: Visualizing the diseasome generated for our case study using RadTrix. (Top row - Leftmost) Overview using the unseriated version. (i) Unary mouse hovering operation to highlight a node to visualize that is common among multiple diseases, (ii) binary mouse hovering operation to highlight `Lung` all 51 genes, (iii) common genes between `Lung` and `GBM`, and (iv) common genes between `Colon` and `Breast`.

ciated with diseases D_i and D_j, for $i, j \in \mathbb{Z}$ and $0 \leq i, j < D$. Similarly, the node size in radial view can be used to visually encode a unary property of the node, e.g., the node degree. In our case study, we use the visual encoding of the size of gene node to indicate 'relative node degree', i.e., the ratio of node degree with the maximum node degree in the diseasome. For easy readability, we use three node sizes, *small*, *medium*, and *large*, to indicate relative node degree less than 0.2, 0.2-0.8, and greater than 0.8, respectively. We can also visually encode the links between gene and disease nodes with their corresponding property, which we have chosen to avoid cognitive overload. We further provide user interactions such as highlight of corresponding node-links and gene nodes on mouse hovering as shown in Figure FC7.6 (i)-(iv), and similarly, highlighting characteristics of unary (Figure FC7.6 (i), (ii)) or binary (Figures FC7.6 (iii), (iv)) properties in the matrix cells, corresponding to the disease and gene nodes.

In addition to readability tasks in [209], the below listed analytical tasks were conducted by RadTrix for diseasome analysis.

- **Task-A:** Analyze the distribution of association across the two sets (of genotypes and phenotypes), namely, to find:

  A1: Distribution of genes corresponding to each disease.

  A2: Distribution of diseases to each gene.

- **Task-B:** Find specific genes exclusive to a specific disease.
- **Task-C**: Find specific genes that belong to at least two diseases.
- **Task-D**: Find the set of diseases a gene associates with.

Using highlighting of links upon mouse-over accomplishes Task-A to Task-D. Mouse-over on the node in the radial layout gives the distribution of association of the corresponding genotype to different diseases. Similarly, mouse-over on the node in the matrix layout gives the distribution of association of the corresponding single or pair of the phenotype(s) with the genotypes. For example, all 51 genes that have been discovered in association with `Lung` cancer can be highlighted as shown in Figure FC7.6 (ii). We have also used seriation to reorder the nodes of the matrix and nodes on the circumference of the radial view. Figure FC7.6 (i)-(iv) shows our results, where we have found two clusters. The cancers `Kidney`, `Breast`, and `Colon` form the first cluster, and `GBM`, `Lung` form the second cluster.

We can conclude that visual representation with RadTrix layout can effectively be used to understand the unbalanced bipartite graphs. The reduced visual clutter due to minimized edge crossings makes the layout more suitable to examine bipartite graphs.

CHAPTER 8

CONCLUSIONS

In this thesis, we have demonstrated novel uses of correlation networks of biomedical data and its clustering to obtain biologically significant conclusions. We have exploited the semantics of correlation networks and used consensus community detection procedures to study biomedical correlation networks. Here, we have presented two different biomedical data analysis problems: i) Finding functional segregation in resting-state functional connectivity network (FCN) by identifying non-overlapping communities and cliques within the communities by applying multiscale consensus procedure using EFA. ii) Finding feature-rich subspace of tumor-specific genes of multi-omics cancer data by utilizing integrative methods, heterogeneous correlation networks, and consensus voting techniques. The identified integrative subspace has been further used to find the patient subtypes in cancer.

8.1 Functional Segregation using Brain FCN

Our motivation is to find salient node-partitioning of the brain resting-state FCN by exploiting the semantics of the correlation matrix that utilizes complete data. Hence, we use a weighted, undirected, and completely connected network for functional segregation. Functional segregation refers to the modules of a network with nodes that are functionally related and tightly connected due to homogeneous edge distribution. We use EFA and consensus method to locate the strongly inter-linked brain ROIs/nodes of the network that are modularity maximized and are biologically significant. Considering the number of factors n_F of EFA as a scale, we implement EFA multiple times with varying values of scale from the predefined interval for n_F, which also addresses the limitations of EFA. The node-partitioning from each EFA run is then aggregated using consensus voting. We identify the most optimal set of multiple scales, i.e., sub-interval, using relevant efficiency metrics followed by a data-driven ranking procedure on the ensemble. The optimal scale is a continuous sub-interval that maximizes the efficiency of the node-partitioning. Our multiscale EFA algorithm is used for finding consensus communities and cliques.

We experiment with our multiscale consensus procedure using three case studies of different resolutions, i.e., node dimensions of FCN and parcellation methods of fMRI data. Our results of communities and cliques are found to be biologically significant such as bilateral symmetry, hierarchy, and persistence of community behavior of the nodes, and found significant based on relevant prior studies. We have witnessed the same using different visualizations such as matrix visualization, Sankey plots/alluvial diagram, and plotting the communities and cliques on brain surface using spatial centroid coordinates. Our proposed algorithm is scalable to the size of the FCN and is generalizable to different parcellation methods. In this work, we use a completely connected network and address the limitation of EFA by exploiting the consensus method. However, in the context of the newer trends in the FCN studies, our study falls short of any analysis based on specific cognitive/task-based studies. At the same time, our study is valuable as a first step towards the comparative analysis of node-partitioning in FCNs for task-based studies. Overall, our method shows how a conventional correlation analysis, namely EFA, can be effectively used with network-based approaches for studying the modular organization of the resting-state brain.

Future work: In our work, the identified sub-interval n_{SI} by our efficiency metrics and data-driven ranking procedure is a continuous interval. Where $n_{SI} = \binom{ub-lb+1}{2}$, for the lower and upper bounds, lb and ub in the identified interval I. We can, however, extend our work to an exhaustive search for sets of multiple scales that include all possibilities of the scales in I, which would be $2^{(ub-lb+1)}$.

For using EFA, there is a strict requirement of positive definiteness to ensure non-singularity for the correlation matrix of the FCN. While a correlation matrix by definition must be positive-definite. But it is not guaranteed due to the complex pre-processing techniques involved in converting fMRI data to the correlation matrix and aggregating matrices across the subjects. It is yet to be studied how a correlation matrix additionally corrected to be positive definite would work with our algorithm.

In our work, we have exclusively focused on non-overlapping (node-) communities and cliques. The scope for work in the future includes studying the significance of overlapping communities and ranked nodes, such as provincial and hub nodes of the network.

8.2 Integrative Subspace and Patient Subtypes of Multi-omics Caner Data

In our case study, we have used expression traits and methylation features from the TCGA breast and lung cancer projects. We have proposed a multi-level integrative process for finding representative integrative subspace and use it for cancer subtype prediction. Our proposed `HCNM` model, is combined with widely used integrative methods to build a more powerful tool to identify the representative subspace and patient subtypes. The `HCNM` model uses correlations within and across the different omic features. `HCNM` has successfully been used for finding the representative integrative subspace of genes based on gene communities, consensus voting, and network topology in both breast and lung cancer phenotypes. The integrative subspace of multi-omics yielded enriched gene-ontology along with significant gene-disease association scores. We have used our subspace of biologically significant genes and appropriate integrative fusion procedures to predict cancer subtypes. As reduced dimension data given by `HCNM` is less prone to noise, bias, and outliers, the subtypes identified are in agreement with available subtypes annotation of literature. In our work, we have explored the subtypes prediction with multi-omics integrative procedures that are fusion-based methods, SNF and ANF, and a non fusion-based, Bayesian method, iCluster. The fusion-based methods outperformed both in terms of execution time and efficiency.

We have examined our workflow on one of the most widely studied cancer subtypes, i.e., breast cancer data, and one of the least studied cancer subtypes, i.e., lung cancer data. The identified patient subtypes in both the cancer phenotypes agree with previous benchmark studies and exhibited better classification between poor and good survival patients. Overall, our proposed multi-level integration model produces representative subspace and hence dimensionality reduction of omic features, which in turn provided significant tumor-specific patient subtypes.

Future work: The scope of our current study is limited to undirected networks, which can be further improved using pathway information or Bayesian networks. Integrating Bayesian methods more extensively into multi-level integration algorithms will improve outcomes, owing to the combination of network-based and Bayesian methods. For evaluating the overall performance of the method, we need to perform an ablation study, which is in the future scope of this work. The extensibility and scalability of our algorithm to other omic features also require further in-depth study. We have examined our `HCNM` on two cancer phenotypes in this work, but the model is applicable to other phenotypes too. The integrative subspace in genes of various diseases can be used to efficiently construct a diseasome, which is a bipartite graph of associations between diseases and omic features. For example, a diseasome between comorbidities associated with COVID-19 such as lung cancer, ARDS, tuberculosis, etc., and ranked subspace in genes. In the future, our work may be extended to obtain an integrated subspace related to a set of diseases by fine-tuning the gene ranking through consensus voting, and constructing a diseasome.

In summary, correlation analysis has been routinely used for the analysis of biomedical data, and our work on correlation networks demonstrates novel uses of such statistical associations. Consequently, this thesis opens up research problems stemming from incorporating the semantics of correlation networks in the data science algorithms for biomedical applications.

Bibliography

[1] Inderjit S. Jutla Lucas G. S. Jeub, Marya Bazzi and Peter J. Mucha. GenLouvain 2.2: A generalized Louvain method for community detection implemented in MATLAB. *MATLAB package, URL https://github.com/GenLouvain/GenLouvain*, 2011-2019.

[2] Mingrui Xia, Jinhui Wang, and Yong He. BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics. *PloS one*, 8(7):e68910, 2013. doi: https://doi.org/10.1371/journal.pone.0068910.

[3] Bharat B. Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M. Smith, Christian F Beckmann, Jonathan S. Adelstein, Randy L. Buckner, Stan Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010. doi: https://doi.org/10.1073/pnas.0911855107.

[4] Yong He, Jinhui Wang, Liang Wang, Zhang J Chen, Chaogan Yan, Hong Yang, Hehan Tang, Chaozhe Zhu, Qiyong Gong, Yufeng Zang, et al. Uncovering intrinsic modular organization of spontaneous brain activity in humans. *PloS one*, 4(4):e5226, 2009. doi: https://doi.org/10.1371/journal.pone.0005226.

[5] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular

interaction networks. *Genome research*, 13(11):2498–2504, 2003. doi: https://doi.org/10.1101/gr.1239303.

[6] BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011. doi: https://doi.org/10.1152/jn.00338.2011.

[7] Francis Galton. I. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45(273-279):135–145, 1889. doi: https://doi.org/10.1037/11304-039.

[8] Karl Pearson. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187):253–318, 1896. doi: https://doi.org/10.1098/rspl.1895.0058.

[9] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140, 1741.

[10] Horst Sachs, Michael Stiebitz, and Robin J Wilson. An historical note: Euler's Königsberg letters. *Journal of Graph Theory*, 12(1):133–139, 1988. doi: https://doi.org/10.1002/jgt.3190120114.

[11] Albert-László Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013. doi: https://doi.org/10.1098/rsta.2012.0375.

[12] Eva Vargas, Francisco J Esteban, and Signe Altmäe. Computational Approaches in Reproductomics. In *Reproductomics*, pages 347–383. Elsevier, 2018. doi: https://doi.org/10.1016/b978-0-12-812571-7.00019-8.

[13] Naiqian Zhang, Haiyun Wang, Yun Fang, Jun Wang, Xiaoqi Zheng, and X Shirley Liu. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Comput Biol*, 11(9):e1004498, 2015. doi: https://doi.org/10.1371/journal.pcbi.1004498.

[14] Khandakar Tanvir Ahmed, Sunho Park, Qibing Jiang, Yunku Yeu, TaeHyun Hwang, and Wei Zhang. Network-based drug sensitivity prediction. *BMC medical genomics*, 13(11):1–10, 2020. doi: https://doi.org/10.1186/s12920-020-00829-3.

[15] Betül Güvenç Paltun, Hiroshi Mamitsuka, and Samuel Kaski. Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. *Briefings in bioinformatics*, 22(1):346–359, 2021. doi: https://doi.org/10.1093/bib/bbz153.

[16] Ana Casteleiro, María Paz-Zulueta, Paula Parás-Bravo, Laura Ruiz-Azcona, and Miguel Santibañez. Association between advanced maternal age and maternal and neonatal morbidity: a cross-sectional study on a Spanish population. *Plos one*, 14(11):e0225074, 2019. doi: https://doi.org/10.1371/journal.pone.0225074.

[17] Naoko Kozuki, Anne CC Lee, Mariangela F Silveira, Ayesha Sania, Joshua P Vogel, Linda Adair, Fernando Barros, Laura E Caulfield, Parul Christian, Wafaie Fawzi, et al. The associations of parity and maternal age with small-for-gestational-age, preterm, and neonatal and infant mortality: a meta-analysis. *BMC public health*, 13(3):1–10, 2013. doi: https://doi.org/10.1186/1471-2458-13-s3-s2.

[18] Emily Kogan, Kathryn Twyman, Jesse Heap, Dejan Milentijevic, Jennifer H Lin, and Mark Alberts. Assessing stroke severity using electronic health record data: a machine learning approach. *BMC medical informatics and decision making*, 20(1):1–8, 2020. doi: https://doi.org/10.1186/s12911-019-1010-x.

[19] Kenney Ng, Steven R Steinhubl, Christopher deFilippi, Sanjoy Dey, and Walter F Stewart. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. *Circulation: Cardiovascular Quality and Outcomes*, 9(6):649–658, 2016. doi: https://doi.org/10.1161/circoutcomes.116.002797.

[20] Sara Gandini, Edoardo Botteri, Simona Iodice, Mathieu Boniol, Albert B Lowenfels, Patrick Maisonneuve, and Peter Boyle. Tobacco smoking and cancer: a meta-analysis. *International journal of cancer*, 122(1):155–164, 2008. doi: https://doi.org/10.1002/ijc.23033.

[21] David H Phillips, Alan Hewer, Carl N Martin, R Colin Garner, and Maureen M King. Correlation of DNA adduct levels in human lung with cigarette smoking. *Nature*, 336(6201):790–792, 1988. doi: https://doi.org/10.1038/336790a0.

[22] Björn H Junker and Falk Schreiber. *Analysis of biological networks*, volume 2. John Wiley & Sons, 2011. doi: 10.1002/9780470253489.

[23] Mavuto M Mukaka. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71, 2012. PMID: 23638278; PMCID: PMC3576830.

[24] Olaf Sporns. Network attributes for segregation and integration in the human brain. *Current opinion in neurobiology*, 23(2):162–171, 2013. doi: https://doi.org/10.1016/j.conb.2012.11.015.

[25] Olaf Sporns, Giulio Tononi, and Gerald M Edelman. Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cerebral cortex*, 10(2):127–141, 2000. doi: https://doi.org/10.1093/cercor/10.2.127.

[26] K-H Boven. Dynamics of activity in neuronal networks give rise to fast modulations of functional connectivity. *Parallel Processing in Neural System and Computers*, 1989.

[27] Karl J Friston. Functional and effective connectivity in neuroimaging: a synthesis. *Human brain mapping*, 2(1-2):56–78, 1994. doi: https://doi.org/10.1002/hbm.460020107.

[28] David T Jones, Prashanthi Vemuri, Matthew C Murphy, Jeffrey L Gunter, Matthew L Senjem, Mary M Machulda, Scott A Przybelski, Brian E Gregg, Kejal Kantarci, David S Knopman, et al. Non-stationarity in the "resting brain's" modular architecture. *PloS one*, 7(6):e39731, 2012. doi: https://doi.org/10.1016/j.jalz.2012.05.1862.

[29] Aaron Alexander-Bloch, Renaud Lambiotte, Ben Roberts, Jay Giedd, Nitin Gogtay, and Ed Bullmore. The discovery of population differences in network community structure: new methods and applications to brain functional networks in schizophrenia. *Neuroimage*, 59(4):3889–3900, 2012. doi: https://doi.org/10.1016/j.neuroimage.2011.11.035.

[30] Sophie H Bennett, Alastair J Kirby, and Gerald T Finnerty. Rewiring the connectome: Evidence and effects. *Neuroscience & Biobehavioral Reviews*, 88:51–62, 2018. doi: https://doi.org/10.1016/j.neubiorev.2018.03.001.

[31] Ben J Harrison, Carles Soriano-Mas, Jesus Pujol, Hector Ortiz, Marina López-Solà, Rosa Hernández-Ribas, Joan Deus, Pino Alonso, Murat Yücel, Christos Pantelis, et al. Altered corticostriatal functional connectivity in obsessive-compulsive disorder. *Archives of general psychiatry*, 66(11):1189–1200, 2009. doi: https://doi.org/10.1001/archgenpsychiatry.2009.152.

[32] Vinod Menon. Large-scale brain networks and psychopathology: a unifying triple network model. *Trends in cognitive sciences*, 15(10):483–506, 2011. doi: https://doi.org/10.1016/j.tics.2011.08.003.

[33] Alex Fornito, Andrew Zalesky, and Michael Breakspear. The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3):159–172, 2015. doi: https://doi.org/10.1038/nrn3901.

[34] Massimo Filippi, Elisabetta Sarasso, and Federica Agosta. Resting-state functional MRI in Parkinsonian syndromes. *Movement disorders clinical practice*, 6(2):104–117, 2019. doi: https://doi.org/10.1002/mdc3.12730.

[35] Fabrizio De Vico Fallani, Jonas Richiardi, Mario Chavez, and Sophie Achard. Graph analysis of functional brain networks: practical issues in translational neuroscience. *Phil. Trans. R. Soc. B*, 369(1653):20130521, 2014. doi: https://doi.org/10.1098/rstb.2013.0521.

[36] Nicolas Langer, Andreas Pedroni, and Lutz Jäncke. The problem of thresholding in small-world network analysis. *PloS one*, 8(1):e53199, 2013. doi: https://doi.org/10.1371/journal.pone.0053199.

[37] Jinhui Wang, Xindi Wang, Mingrui Xia, Xuhong Liao, Alan Evans, and Yong He. GRETNA: a graph theoretical network analysis toolbox for imaging connectomics. *Frontiers in human neuroscience*, 9:386, 2015. doi: https://doi.org/10.3389/fnhum.2015.00386.

[38] Charles Spearman. General Intelligence, objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904. doi: https://doi.org/10.2307/1412107.

[39] Jason W Osborne and David C Fitzpatrick. Replication analysis in exploratory factor analysis: What it is and why it makes your analysis bet-

ter. *Practical assessment, research, and evaluation*, 17(1):15, 2012. doi: https://doi.org/10.7275/h0bd-4d11.

[40] Kristopher J Preacher, Guangjian Zhang, Cheongtag Kim, and Gerhard Mels. Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, 48(1):28–56, 2013. doi: https://doi.org/10.1080/00273171.2012.710386.

[41] Reddy Rani Vangimalla and Jaya Sreevalsan-Nair. A Multiscale Consensus Method Using Factor Analysis to Extract Modular Regions in the Functional Brain Network. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2824–2828. IEEE, 2020. doi: https://doi.org/10.1109/embc44109.2020.9175622.

[42] Reddy Rani Vangimalla, Hyun-hwan Jeong, and Kyung-Ah Sohn. Integrative regression network for genomic association study. *BMC medical genomics*, 9(1):31, 2016. doi: https://doi.org/10.1186/s12920-016-0192-7.

[43] Mingguang Shi, Junwen Wang, and Chenyu Zhang. Integration of Cancer Genomics Data for Tree-based Dimensionality Reduction and Cancer Outcome Prediction. *Molecular Informatics*, 39(3):1900028, 2020. doi: https://doi.org/10.1002/minf.201900028.

[44] So Yeon Kim, Hyun-Hwan Jeong, Jaesik Kim, Jeong-Hyeon Moon, and Kyung-Ah Sohn. Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies. *Biology direct*, 14(1):1–13, 2019. doi: https://doi.org/10.1186/s13062-019-0239-8.

[45] Bujun Mei and Zhihua Wang. An efficient method to handle the ‘large p, small n’problem for genomewide association studies using Haseman–Elston regression. *Journal of genetics*, 95(4):847–852, 2016. doi: https://doi.org/10.1007/s12041-016-0705-3.

[46] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, 17(2):167–177, 2016. doi: https://doi.org/10.1186/s12859-015-0857-9.

[47] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014. doi: https://doi.org/10.1093/comnet/cnu016.

[48] Olaf Sporns, Giulio Tononi, and Rolf Kötter. The human connectome: a structural description of the human brain. *PLoS Comput Biol*, 1(4):e42, 2005. doi: https://doi.org/10.1371/journal.pcbi.0010042.

[49] Peter J Basser, James Mattiello, and Denis LeBihan. MR diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267, 1994. doi: https://doi.org/10.1016/s0006-3495(94)80775-1.

[50] Patric Hagmann, Maciej Kurant, Xavier Gigandet, Patrick Thiran, Van J Wedeen, Reto Meuli, and Jean-Philippe Thiran. Mapping human whole-brain structural networks with diffusion MRI. *PloS one*, 2(7):e597, 2007. doi: https://doi.org/10.1371/journal.pone.0000597.

[51] Christopher J Honey, Olaf Sporns, Leila Cammoun, Xavier Gigandet, Jean-Philippe Thiran, Reto Meuli, and Patric Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040, 2009. doi: https://doi.org/10.1073/pnas.0811168106.

[52] Matthew Lawrence Stanley, Malaak Nasser Moussa, Brielle Paolini, Robert Gray Lyday, Jonathan H Burdette, and Paul J Laurienti. Defining nodes in complex brain networks. *Frontiers in computational neuroscience*, 7:169, 2013. doi: https://doi.org/10.3389/fncom.2013.00169.

[53] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fMRI functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010. doi: https://doi.org/10.1016/j.euroneuro.2010.03.008.

[54] Hae-Jeong Park and Karl Friston. Structural and functional brain networks: from connections to cognition. *Science*, 342(6158), 2013. doi: https://doi.org/10.1126/science.1238411.

[55] Fabrizio de Vico Fallani, Jonas Richiardi, Mario Chavez, and Sophie Achard. Graph analysis of functional brain networks: practical issues in translational neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1653):20130521, 2014. doi: https://doi.org/10.1098/rstb.2013.0521.

[56] Shuai Huang, Jing Li, Jieping Ye, Adam Fleisher, Kewei Chen, Teresa Wu, and Eric Reiman. Brain effective connectivity modeling for Alzheimer's disease by sparse Gaussian Bayesian network. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 931–939, 2011. doi: https://doi.org/10.1145/2020408.2020562.

[57] MBCFDS De Luca, Christian F Beckmann, Nicola De Stefano, Paul M Matthews, and Stephen M Smith. fMRI resting state networks define distinct modes of long-distance interactions in the human brain. *Neuroimage*, 29(4):1359–1367, 2006. doi: https://doi.org/10.1016/j.neuroimage.2005.08.035.

[58] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289, 2002. doi: https://doi.org/10.1006/nimg.2001.0978.

[59] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2017. doi: https://doi.org/10.1093/cercor/bhx179.

[60] Korbinian Brodmann. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth, 1909.

[61] David Meunier, Renaud Lambiotte, Alex Fornito, Karen Ersche, and Edward T Bullmore. Hierarchical modularity in human brain functional networks. *Frontiers in neuroinformatics*, 3:37, 2009. doi: https://doi.org/10.3389/neuro.11.037.2009.

[62] Onerva Korhonen, Heini Saarimäki, Enrico Glerean, Mikko Sams, and Jari Saramäki. Consistency of regions of interest as nodes of fMRI functional brain networks. *Network Neuroscience*, 1(3):254–274, 2017. doi: https://doi.org/10.1162/netn_a_00013.

[63] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010. doi: https://doi.org/10.1016/j.neuroimage.2009.10.003.

[64] Olaf Sporns. Networks of the brain: quantitative analysis and modeling. *Analysis and function of large-scale brain networks*, 7:7–13, 2010.

[65] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’networks. *nature*, 393(6684):440–442, 1998. doi: https://doi.org/10.1038/30918.

[66] Xia Liang, Jinhui Wang, Chaogan Yan, Ni Shu, Ke Xu, Gaolang Gong, and Yong He. Effects of different correlation metrics and preprocessing factors on small-

world brain functional networks: a resting-state functional MRI study. *PloS one*, 7(3):e32766, 2012. doi: https://doi.org/10.1371/journal.pone.0032766.

[67] Martijn P van den Heuvel, Cornelis J Stam, Maria Boersma, and HE Hulshoff Pol. Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain. *Neuroimage*, 43(3):528–539, 2008. doi: https://doi.org/10.1016/j.neuroimage.2008.08.010.

[68] Megan H. Lee, Carl D Hacker, Abraham Z. Snyder, Maurizio Corbetta, Dongyang Zhang, Eric C. Leuthardt, and Joshua S. Shimony. Clustering of resting state networks. *PloS one*, 7(7):e40370, 2012. doi: https://doi.org/10.1371/journal.pone.0040370.

[69] Richard F Betzel, Alessandra Griffa, Andrea Avena-Koenigsberger, Joaquín Goñi, Jean-Philippe Thiran, Patric Hagmann, and Olaf Sporns. Multi-scale community organization of the human structural connectome and its relationship with resting-state functional connectivity. *Network Science*, 1(3):353–373, 2013. doi: https://doi.org/10.1017/nws.2013.19.

[70] Nico UF Dosenbach, Binyam Nardos, Alexander L Cohen, Damien A Fair, Jonathan D Power, Jessica A Church, Steven M Nelson, Gagan S Wig, Alecia C Vogel, Christina N Lessov-Schlaggar, et al. Prediction of individual brain maturity using fMRI. *Science*, 329(5997):1358–1361, 2010. doi: https://doi.org/10.1126/science.1194144.

[71] Salim Arslan, Sofia Ira Ktena, Antonios Makropoulos, Emma C Robinson, Daniel Rueckert, and Sarah Parisot. Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex. *NeuroImage*, 170:5–30, 2018. doi: https://doi.org/10.1016/j.neuroimage.2017.04.014.

[72] Vangelis Sakkalis. Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Com-*

puters in biology and medicine, 41(12):1110–1117, 2011. doi: https://doi.org/10.1016/j.compbiomed.2011.06.020.

[73] Jonathan D. Power, Alexander L. Cohen, Steven M. Nelson, Gagan S. Wig, Kelly Anne Barnes, Jessica A. Church, Alecia C. Vogel, Timothy O. Laumann, Fran M. Miezin, Bradley L. Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011. doi: https://doi.org/10.1016/j.neuron.2011.09.006.

[74] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008. doi: https://doi.org/10.1073/pnas.0706851105.

[75] Guillaume Marrelec, Alexandre Krainik, Hugues Duffau, Mélanie Pélégrini-Issac, Stéphane Lehéricy, Julien Doyon, and Habib Benali. Partial correlation for functional brain interactivity investigation in functional MRI. *Neuroimage*, 32(1):228–237, 2006. doi: https://doi.org/10.1016/j.neuroimage.2005.12.057.

[76] Felice T Sun, Lee M Miller, and Mark D'Esposito. Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data. *Neuroimage*, 21(2):647–658, 2004. doi: https://doi.org/10.1016/j.neuroimage.2003.09.056.

[77] Barry Chai, Dirk B Walther, Diane M Beck, and Li Fei-Fei. Exploring functional connectivity of the human brain using multivariate information analysis. *Advances in neural information processing systems*, 22:270–278, 2009.

[78] Gopikrishna Deshpande, Stephan LaConte, George Andrew James, Scott Peltier, and Xiaoping Hu. Multivariate Granger causality analysis of fMRI data. *Human brain mapping*, 30(4):1361–1373, 2009. doi: https://doi.org/10.1002/hbm.20606.

[79] Olaf Sporns and Richard F Betzel. Modular brain networks. *Annual review of psychology*, 67:613–640, 2016. doi: https://doi.org/10.1146/annurev-psych-122414-033634.

[80] R Matthew Hutchison, Thilo Womelsdorf, Elena A Allen, Peter A Bandettini, Vince D Calhoun, Maurizio Corbetta, Stefania Della Penna, Jeff H Duyn, Gary H Glover, Javier Gonzalez-Castillo, et al. Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage*, 80:360–378, 2013. doi: https://doi.org/10.1016/j.neuroimage.2013.05.079.

[81] World Health Organization et al. *Genomics and world health: Report of the Advisory Committee on Health Research*. World Health Organization, 2002.

[82] International Cancer Genome Consortium et al. International network of cancer genome projects. *Nature*, 464(7291):993, 2010. doi: https://doi.org/10.1038/nature08987.

[83] ICGC The, TCGA Pan-Cancer Analysis of Whole, Genomes Consortium, et al. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82, 2020. doi: https://doi.org/10.1038/s41586-020-1969-6.

[84] Peter A Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, 2012. doi: https://doi.org/10.1038/nrg3230.

[85] Ning Xu, Yu-Peng Wu, Zhi-Bin Ke, Ying-Chun Liang, Hai Cai, Wen-Ting Su, Xuan Tao, Shao-Hao Chen, Qing-Shui Zheng, Yong Wei, et al. Identification of key DNA methylation-driven genes in prostate adenocarcinoma: an integrative analysis of TCGA methylation data. *Journal of translational medicine*, 17(1):1–15, 2019. doi: https://doi.org/10.1186/s12967-019-2065-2.

[86] John CG Spainhour, Hong Seo Lim, Soojin V Yi, and Peng Qiu. Correlation patterns between DNA methylation and gene expression in The Can-

cer Genome Atlas. *Cancer informatics*, 18:1176935119828776, 2019. doi: https://doi.org/10.1177/1176935119828776.

[87] Zhiming Li, Xuan Zhuang, Jinxiong Zeng, and Chi-Meng Tzeng. Integrated analysis of DNA methylation and mRNA expression profiles to identify key genes in severe oligozoospermia. *Frontiers in physiology*, 8:261, 2017. doi: https://doi.org/10.3389/fphys.2017.00261.

[88] Brittany Baur and Serdar Bozdag. A feature selection algorithm to compute gene centric methylation from probe level methylation data. *PloS one*, 11(2):e0148977, 2016. doi: https://doi.org/10.1371/journal.pone.0148977.

[89] Chao Chen, Chunling Zhang, Lijun Cheng, James L Reilly, Jeffrey R Bishop, John A Sweeney, Hua-Yun Chen, Elliot S Gershon, and Chunyu Liu. Correlation between DNA methylation and gene expression in the brains of patients with bipolar disorder and schizophrenia. *Bipolar disorders*, 16(8):790–799, 2014. doi: https://doi.org/10.1111/bdi.12255.

[90] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008. doi: https://doi.org/10.1186/1471-2105-9-559.

[91] Sipko Van Dam, Urmo Vosa, Adriaan van der Graaf, Lude Franke, and Joao Pedro de Magalhaes. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19(4):575–592, 2018. doi: https://doi.org/10.1093/bib/bbw139.

[92] Baoling Liu, Guanhong Huang, Hongming Zhu, Zhaoming Ma, Xiaokang Tian, Li Yin, Xingya Gao, and Xia He. Analysis of gene co-expression network reveals prognostic significance of CNFN in patients with head and neck cancer. *Oncology reports*, 41(4):2168–2180, 2019. doi: https://doi.org/10.3892/or.2019.7019.

[93] Juliane Schäfer and Korbinian Strimmer. Learning Large-Scale Graphical Gaussian Models from Genomic Data. In *AIP Conference Proceedings*, volume 776, pages 263–276. American Institute of Physics, 2005. doi: https://doi.org/10.1063/1.1985393.

[94] Juliane Schäfer and Korbinian Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005. doi: https://doi.org/10.1093/bioinformatics/bti062.

[95] Åsa Johansson, Mari Løset, Siv B Mundal, Matthew P Johnson, Katy A Freed, Mona H Fenstad, Eric K Moses, Rigmor Austgulen, and John Blangero. Partial correlation network analyses to detect altered gene interactions in human disease: using preeclampsia as a model. *Human genetics*, 129(1):25–34, 2011. doi: https://doi.org/10.1007/s00439-010-0893-5.

[96] Aziz M Mezlini, Bo Wang, Amit Deshwar, Quaid Morris, and Anna Goldenberg. Identifying cancer specific functionally relevant miRNAs from gene expression and miRNA-to-gene networks using regularized regression. *PloS one*, 8(10):e73168, 2013. doi: https://doi.org/10.1371/journal.pone.0073168.

[97] Ander Muniategui, Rubén Nogales-Cadenas, Miguél Vázquez, Xabier L Aranguren, Xabier Agirre, Aernout Luttun, Felipe Prosper, Alberto Pascual-Montano, and Angel Rubio. Quantification of miRNA-mRNA interactions. *PloS one*, 7(2):e30766, 2012. doi: https://doi.org/10.1371/journal.pone.0030766.

[98] Maarten van Iterson, Sander Bervoets, Emile J de Meijer, Henk P Buermans, Peter AC ’t Hoen, Renee X Menezes, and Judith M Boer. Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions. *Nucleic acids research*, 41(15):e146–e146, 2013. doi: https://doi.org/10.1093/nar/gkt525.

[99] Sanghoon Lee and Xia Jiang. Modeling miRNA-mRNA interactions that cause phenotypic abnormality in breast cancer patients. *PloS one*, 12(8):e0182666, 2017. doi: https://doi.org/10.1371/journal.pone.0182666.

[100] Kyung-Ah Sohn, Dokyoon Kim, Jaehyun Lim, and Ju Han Kim. Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors. *BMC systems biology*, 7(S6):S9, 2013. doi: https://doi.org/10.1186/1752-0509-7-s6-s9.

[101] Huiying Qi and Yuhe Jiang. Predicting Breast Cancer Survival Length with Multi-Omics Data Fusion. *Data Analysis and Knowledge Discovery*, 3(8):88–93, 2019. doi: 10.11925/infotech.2096-3467.2019.0021.

[102] Dokyoon Kim, Hyunjung Shin, Young Soo Song, and Ju Han Kim. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *Journal of biomedical informatics*, 45(6):1191–1198, 2012. doi: https://doi.org/10.1016/j.jbi.2012.07.008.

[103] Hao Ding, Chao Wang, Kun Huang, and Raghu Machiraju. iGPSe: a visual analytic system for integrative genomic based cancer patient stratification. *BMC bioinformatics*, 15(1):203, 2014. doi: https://doi.org/10.1186/1471-2105-15-203.

[104] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56, 2011. doi: https://doi.org/10.1038/nrg2918.

[105] Marko Gosak, Rene Markovič, Jurij Dolenšek, Marjan Slak Rupnik, Marko Marhl, Andraž Stožer, and Matjaž Perc. Network science of biological systems at different scales: a review. *Physics of life reviews*, 24:118–135, 2018. doi: https://doi.org/10.1016/j.plrev.2017.11.003.

[106] M De Domenico. Multilayer network modeling of integrated biological systems: Comment on" Network science of biological systems at different scales: A review" by Gosak et al. *Physics of life reviews*, 24:149, 2018. doi: https://doi.org/10.1016/j.plrev.2017.12.006.

[107] John G White, Eileen Southgate, J Nichol Thomson, and Sydney Brenner. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci*, 314(1165):1–340, 1986.

[108] Yunkyu Sohn, Myung-Kyu Choi, Yong-Yeol Ahn, Junho Lee, and Jaeseung Jeong. Topological cluster analysis reveals the systemic organization of the Caenorhabditis elegans connectome. *PLoS Comput Biol*, 7(5):e1001139, 2011. doi: https://doi.org/10.1371/journal.pcbi.1001139.

[109] Dragana M Pavlovic, Petra E Vértes, Edward T Bullmore, William R Schafer, and Thomas E Nichols. Stochastic blockmodeling of the modules and core of the Caenorhabditis elegans connectome. *PloS one*, 9(7):e97584, 2014. doi: https://doi.org/10.1371/journal.pone.0097584.

[110] Reddy Rani Vangimalla and Jaya Sreevalsan-Nair. Comparing Community Detection Methods in Brain Functional Connectivity Networks. In *International Conference on Computational Intelligence, Cyber Security, and Computational Models*, pages 3–17. Springer, 2019. doi: https://doi.org/10.1007/978-981-15-9700-8_1.

[111] Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, Ronald L Tatham, et al. *Multivariate data analysis*, volume 5. Prentice hall Upper Saddle River, NJ, 1998.

[112] Rudolph J Rummel. Understanding factor analysis. *Journal of conflict resolution*, 11(4):444–480, 1967. doi: https://doi.org/10.1177/002200276701100405.

[113] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010. doi: https://doi.org/10.1016/j.physrep.2009.11.002.

[114] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198, 2009. doi: https://doi.org/10.1038/nrn2575.

[115] Cécile Bordier, Carlo Nicolini, and Angelo Bifone. Graph analysis and modularity of brain functional connectivity networks: searching for the optimal threshold. *Frontiers in neuroscience*, 11:441, 2017. doi: https://doi.org/10.3389/fnins.2017.00441.

[116] Matt C. Howard. A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, 32(1):51–62, 2016. doi: https://doi.org/10.1080/10447318.2015.1087664.

[117] Alvin C. Rencher. *Methods of multivariate analysis*, volume 492. John Wiley & Sons, 2003. ISBN 0-471-41889-7.

[118] Anna B. Costello and Jason W. Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation*, 10(7):1–9, 2005. doi: https://doi.org/10.7275/jyj1-4868.

[119] Jason W Osborne, Anna B Costello, and J Thomas Kellow. *Best practices in exploratory factor analysis*. CreateSpace Independent Publishing Platform Louisville, KY, 2014. doi: https://doi.org/10.4135/9781412995627.d8.

[120] James Dean Brown. Choosing the right type of rotation in PCA and EFA. *JALT testing & evaluation SIG newsletter*, 13(3):20–25, 2009.

[121] Henry F Kaiser and John Rice. Little jiffy, mark IV. *Educational and psychological measurement*, 34(1):111–117, 1974. doi: https://doi.org/10.1177/001316447403400115.

[122] Henry F Kaiser. A second generation little jiffy. *Psychometrika*, 35(4):401–415, 1970. doi: https://doi.org/10.1007/bf02291817.

[123] Rubén Daniel Ledesma and Pedro Valero-Mora. Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical assessment, research, and evaluation*, 12(1):2, 2007. doi: https://doi.org/10.7275/wjnc-nm63.

[124] Dennis Child. *The essentials of factor analysis*. A&C Black, 2006.

[125] Günce Orman, Vincent Labatut, and Hocine Cherifi. On Accuracy of Community Structure Discovery Algorithms. *Journal of Convergence Information Technology*, 6(11):283–292, 2011. doi: https://doi.org/10.4156/jcit.vol6.issue11.32.

[126] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. doi: https://doi.org/10.1088/1742-5468/2008/10/p10008.

[127] Christian Lohse, Danielle S. Bassett, Kelvin O. Lim, and Jean M. Carlson. Resolving anatomical and functional structure in human brain organization: identifying mesoscale organization in weighted network representations. *PLoS computational biology*, 10(10):e1003712, 2014. doi: https://doi.org/10.1371/journal.pcbi.1003712.

[128] Inderjit S Jutla, Lucas GS Jeub, and Peter J Mucha. A generalized Louvain method for community detection implemented in MATLAB. *URL - http://netwiki. amath. unc. edu/GenLouvain*, 2011.

[129] Lucas GS Jeub, Olaf Sporns, and Santo Fortunato. Hierarchical Consensus clustering implemented in MATLAB. *URL - https://github.com/LJeub/HierarchicalConsensus*, 2018.

[130] Khairi Reda, Chayant Tantipathananandh, Andrew Johnson, Jason Leigh, and Tanya Berger-Wolf. Visualizing the evolution of community structures in dynamic social networks. *Computer Graphics Forum*, 30(3):1061–1070, 2011. doi: https://doi.org/10.1111/j.1467-8659.2011.01955.x.

[131] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004. doi: https://doi.org/10.1103/physreve.69.026113.

[132] Mark E. J. Newman. Analysis of weighted networks. *Physical review E*, 70(5):056131, 2004. doi: https://doi.org/10.1103/PhysRevE.70.056131.

[133] Mark EJ Newman. Mixing patterns in networks. *Physical review E*, 67(2):026126, 2003. doi: https://doi.org/10.1103/physreve.67.026126.

[134] Andrea Lancichinetti and Santo Fortunato. Consensus clustering in complex networks. *Scientific reports*, 2:336, 2012. doi: https://doi.org/10.1038/srep00336.

[135] Lucas GS Jeub, Olaf Sporns, and Santo Fortunato. Multiresolution consensus clustering in networks. *Scientific reports*, 8(1):3259, 2018. doi: https://doi.org/10.1038/s41598-018-21352-7.

[136] LNF Ana and Anil K Jain. Robust data clustering. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–II. IEEE, 2003. doi: 10.1109/CVPR.2003.1211462.

[137] David C Van Essen, Matthew F Glasser, Donna L Dierker, John Harwell, and Timothy Coalson. Parcellations and hemispheric asymmetries of human cerebral

cortex analyzed on surface-based atlases. *Cerebral cortex*, 22(10):2241–2262, 2012. doi: https://doi.org/10.1093/cercor/bhr291.

[138] Kate Brody Nooner, Stanley Colcombe, Russell Tobe, Maarten Mennes, Melissa Benedict, Alexis Moreno, Laura Panek, Shaquanna Brown, Stephen Zavitz, Qingyang Li, et al. The NKI-Rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Frontiers in neuroscience*, 6:152, 2012. doi: https://doi.org/10.3389/fnins.2012.00152.

[139] Steen Moeller, Essa Yacoub, Cheryl A Olman, Edward Auerbach, John Strupp, Noam Harel, and Kâmil Uğurbil. Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic resonance in medicine*, 63(5):1144–1153, 2010. doi: https://doi.org/10.1002/mrm.22361.

[140] David A Feinberg, Steen Moeller, Stephen M Smith, Edward Auerbach, Sudhir Ramanna, Matt F Glasser, Karla L Miller, Kamil Ugurbil, and Essa Yacoub. Multiplexed echo planar imaging for sub-second whole brain FMRI and fast diffusion imaging. *PloS one*, 5(12):e15710, 2010. doi: https://doi.org/10.1371/journal.pone.0015710.

[141] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature methods*, 16(1):111–116, 2019. doi: https://doi.org/10.1038/s41592-018-0235-4.

[142] Theodore D Satterthwaite, Mark A Elliott, Raphael T Gerraty, Kosha Ruparel, James Loughead, Monica E Calkins, Simon B Eickhoff, Hakon Hakonarson, Ruben C Gur, Raquel E Gur, et al. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of

resting-state functional connectivity data. *Neuroimage*, 64:240–256, 2013. doi: https://doi.org/10.1016/j.neuroimage.2012.08.052.

[143] Bruce Fischl, André Van Der Kouwe, Christophe Destrieux, Eric Halgren, Florent Ségonne, et al. Automatically parcellating the human cerebral cortex. *Cerebral cortex*, 14(1):11–22, 2004. doi: https://doi.org/10.1093/cercor/bhg087.

[144] Micaela Y Chan, Denise C Park, Neil K Savalia, Steven E Petersen, and Gagan S Wig. Decreased segregation of brain systems across the healthy adult lifespan. *Proceedings of the National Academy of Sciences*, 111(46):E4997–E5006, 2014. doi: https://doi.org/10.1073/pnas.1415122111.

[145] Richard F Betzel, Lisa Byrge, Ye He, Joaquín Goñi, Xi-Nian Zuo, and Olaf Sporns. Changes in structural and functional connectivity among resting-state networks across the human lifespan. *Neuroimage*, 102:345–357, 2014. doi: https://doi.org/10.1016/j.neuroimage.2014.07.067.

[146] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. doi: https://doi.org/10.1016/0005-1098(78)90005-5.

[147] Christian F. Beckmann and Stephen M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging*, 23(2):137–152, 2004. doi: https://doi.org/10.1109/tmi.2003.822821.

[148] Leandre R. Fabrigar, Duane T. Wegener, Robert C. MacCallum, and Erin J Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3):272, 1999. doi: https://doi.org/10.1037/1082-989x.4.3.272.

[149] Feinian Chen, Patrick J Curran, Kenneth A Bollen, James Kirby, and Pamela Paxton. An empirical evaluation of the use of fixed cutoff points in RMSEA

test statistic in structural equation models. *Sociological methods & research*, 36(4):462–494, 2008. doi: https://doi.org/10.1177/0049124108314720.

[150] Ana Fred. Finding consistent clusters in data partitions. In *International Workshop on Multiple Classifier Systems*, pages 309–318. Springer, 2001. doi: https://doi.org/10.1007/3-540-48219-9_31.

[151] Ana LN Fred and Anil K Jain. Data clustering using evidence accumulation. In *Object recognition supported by user interaction for service robots*, volume 4, pages 276–280. IEEE, 2002. doi: https://doi.org/10.1109/icpr.2002.1047450.

[152] Ana LN Fred and Anil K Jain. Combining multiple clusterings using evidence accumulation. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):835–850, 2005. doi: https://doi.org/10.1109/tpami.2005.113.

[153] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical review E*, 74(1):016110, 2006. doi: https://doi.org/10.1103/PhysRevE.74.016110.

[154] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[155] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[156] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. doi: https://doi.org/10.1016/0377-0427(87)90125-7.

[157] Joseph C Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. 1973. doi: https://doi.org/10.1080/01969727308546046.

[158] Innar Liiv. Seriation and matrix reordering methods: A historical overview. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 3(2):70–91, 2010. doi: https://doi.org/10.1002/sam.10071.

[159] Wei Liao, Jurong Ding, Daniele Marinazzo, Qiang Xu, Zhengge Wang, Cuiping Yuan, Zhiqiang Zhang, Guangming Lu, and Huafu Chen. Small-world directed networks in the human brain: multivariate Granger causality analysis of resting-state fMRI. *Neuroimage*, 54(4):2683–2694, 2011. doi: https://doi.org/10.1016/j.neuroimage.2010.11.007.

[160] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*, 19(1A):A68, 2015. doi: https://doi.org/10.5114/wo.2014.47136.

[161] Shu-Guang Ge, Junfeng Xia, Wen Sha, and Chun-Hou Zheng. Cancer subtype discovery based on integrative model of multigenomic data. *IEEE/ACM trans. on comput. bio. & bioin. (TCBB)*, 14(5):1115–1121, 2016. doi: https://doi.org/10.1109/tcbb.2016.2621769.

[162] Chen Peng, Ao Li, and Minghui Wang. Discovery of Bladder Cancer-related Genes Using Integrative Heterogeneous Network Modeling of Multi-omics Data. *Scientific reports*, 7(1):1–11, 2017. doi: https://doi.org/10.1038/s41598-017-15890-9.

[163] Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, Isabella Castiglioni, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research*, 44(8):e71–e71, 2016. doi: https://doi.org/10.1093/nar/gkv1507.

[164] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fu-

sion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333, 2014. doi: https://doi.org/10.1038/nmeth.2810.

[165] Dokyoon Kim, Ruowang Li, Scott M Dudek, and Marylyn D Ritchie. Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer. *Journal of biomedical informatics*, 56:220–228, 2015. doi: https://doi.org/10.1016/j.jbi.2015.05.019.

[166] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: https://doi.org/10.1111/j.1467-9868.2011.00771.x.

[167] Alberto De La Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004. doi: https://doi.org/10.1093/bioinformatics/bth445.

[168] Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic press, 2011. doi: https://doi.org/10.1016/C2010-0-67044-1.

[169] David A Lax. Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, 80(391):736–741, 1985. doi: https://doi.org/10.1080/01621459.1985.10478177.

[170] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998. doi: https://doi.org/10.1073/pnas.95.25.14863.

[171] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pages 284–293. Springer, 2005. doi: https://doi.org/10.7155/jgaa.00124.

[172] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004. doi: https://doi.org/10.1103/physreve.70.066111.

[173] Cecily J Wolfe, Isaac S Kohane, and Atul J Butte. Systematic survey reveals general applicability of" guilt-by-association" within gene coexpression networks. *BMC bioinformatics*, 6(1):227, 2005. doi: https://doi.org/10.1186/1471-2105-6-227.

[174] Bhavesh R Borate, Elissa J Chesler, Michael A Langston, Arnold M Saxton, and Brynn H Voy. Comparison of threshold selection methods for microarray gene co-expression matrices. *BMC research notes*, 2(1):240, 2009. doi: https://doi.org/10.1186/1756-0500-2-240.

[175] Andy D Perkins and Michael A Langston. Threshold selection in gene co-expression networks using spectral graph theory techniques. In *BMC bioinformatics*, volume 10, page S4. BioMed Central, 2009. doi: https://doi.org/10.1186/1471-2105-10-s11-s4.

[176] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963. doi: https://doi.org/10.1080/01621459.1963.10500845.

[177] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004. doi: https://doi.org/10.1103/physreve.69.066133.

[178] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980):876–878, 2010. doi: https://doi.org/10.1126/science.1184819.

[179] Charles Spearman. The proof and measurement of association between two things. 1961. doi: https://doi.org/10.1037/11491-005.

[180] Jan Hauke and Tomasz Kossowski. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93, 2011. doi: https://doi.org/10.2478/v10117-011-0021-1.

[181] Brad T Sherman, Richard A Lempicki, et al. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44, 2009. doi: https://doi.org/10.1038/nprot.2008.211.

[182] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2009. doi: https://doi.org/10.1093/nar/gkn923.

[183] Ashton C Berger, Anil Korkut, Rupa S Kanchi, Apurva M Hegde, Walter Lenoir, Wenbin Liu, Yuexin Liu, Huihui Fan, Hui Shen, Visweswaran Ravikumar, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer cell*, 33(4):690–705, 2018. doi: https://doi.org/10.1016/j.ccell.2018.03.014.

[184] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012. doi: https://doi.org/10.1038/nature11412.

[185] Cancer Genome Atlas Research Network et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519, 2012. doi: https://doi.org/10.1038/nature11404.

[186] Tianle Ma and Aidong Zhang. Integrate multi-omic data using affinity network fusion (ANF) for cancer patient clustering. In *2017 IEEE International Confer-*

ence on Bioinformatics and Biomedicine (BIBM), pages 398–403. IEEE, 2017. doi: https://doi.org/10.1109/bibm.2017.8217682.

[187] Marco Chierici, Nicole Bussola, Alessia Marcolini, Margherita Francescatto, Alessandro Zandonà, Lucia Trastulla, Claudio Agostinelli, Giuseppe Jurman, and Cesare Furlanello. Integrative Network Fusion: a multi-omics approach in molecular profiling. *Frontiers in oncology*, 10:1065, 2020. doi: https://doi.org/10.3389/fonc.2020.01065.

[188] Ronglai Shen, Qianxing Mo, Nikolaus Schultz, Venkatraman E Seshan, Adam B Olshen, Jason Huse, Marc Ladanyi, and Chris Sander. Integrative subtype discovery in glioblastoma using iCluster. *PloS one*, 7(4):e35236, 2012. doi: https://doi.org/10.1371/journal.pone.0035236.

[189] Tianle Ma and Aidong Zhang. Affinity network fusion and semi-supervised learning for cancer patient clustering. *Methods*, 145:16–24, 2018. doi: https://doi.org/10.1016/j.ymeth.2018.05.020.

[190] Qianxing Mo and Ronglai Shen. Package ‘iClusterPlus’. 2018. URL: https://bioconductor.org/packages/release/bioc/html/iClusterPlus.html.

[191] Jacques Bertin. *Graphics and graphic information processing*. Walter de Gruyter, 2011. doi: https://doi.org/10.1515/9783110854688.

[192] James C Bezdek, Richard J Hathaway, and Jacalyn M Huband. Visual assessment of clustering tendency for rectangular dissimilarity matrices. *IEEE Transactions on fuzzy systems*, 15(5):890–903, 2007. doi: https://doi.org/10.1109/tfuzz.2006.889956.

[193] Chun-Houh Chen. Generalized association plots: Information visualization via iteratively generated correlation matrices. *Statistica Sinica*, pages 7–29, 2002. doi: https://www.jstor.org/stable/24307033.

[194] F Marcotorchino. Seriation problems: an overview. *Applied stochastic models and Data Analysis*, 7(2):139–151, 1991. doi: https://doi.org/10.1002/asm.3150070204.

[195] Gilles Caraux and Sylvie Pinloche. PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics*, 21(7):1280–1281, 2005. doi: https://doi.org/10.1093/bioinformatics/bti141.

[196] Urvakhsh Meherwan Mehta, Darshan Shadakshari, Pulaparambil Vani, Shalini S Naik, V Kiran Raj, Reddy Rani Vangimalla, YC Janardhan Reddy, Jaya Sreevalsan-Nair, and Rose Dawn Bharath. Case Report: Obsessive compulsive disorder in posterior cerebellar infarction-illustrating clinical and functional connectivity modulation using MRI-informed transcranial magnetic stimulation. *Wellcome Open Research*, 5, 2020. doi: https://doi.org/10.12688/wellcomeopenres.16183.2.

[197] Mark W Woolrich, Saad Jbabdi, Brian Patenaude, Michael Chappell, Salima Makni, Timothy Behrens, Christian Beckmann, Mark Jenkinson, and Stephen M Smith. Bayesian analysis of neuroimaging data in FSL. *Neuroimage*, 45(1):S173–S186, 2009. doi: https://doi.org/10.1016/j.neuroimage.2008.10.055.

[198] Jean A Frazier, Sufen Chiu, Janis L Breeze, Nikos Makris, Nicholas Lange, David N Kennedy, Martha R Herbert, Eileen K Bent, Vamsi K Koneru, Megan E Dieterich, et al. Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *American Journal of Psychiatry*, 162(7):1256–1265, 2005. doi: https://doi.org/10.1176/appi.ajp.162.7.1256.

[199] Jörn Diedrichsen, Joshua H Balsters, Jonathan Flavell, Emma Cussans, and Narender Ramnani. A probabilistic MR atlas of the human cerebellum. *Neuroimage*, 46(1):39–46, 2009. doi: https://doi.org/10.1016/s1053-8119(09)71166-8.

[200] William S Robinson. A method for chronologically ordering archaeological deposits. *American antiquity*, 16(4):293–301, 1951. doi: https://doi.org/10.2307/276978.

[201] Michael Hahsler, Kurt Hornik, and Christian Buchta. Getting things in order: an introduction to the R package seriation. *Journal of Statistical Software*, 25(3):1–34, 2008. doi: https://doi.org/10.18637/jss.v025.i03.

[202] John Listerud, Chivon Powers, Peachie Moore, David J Libon, and Murray Grossman. Neuropsychological patterns in magnetic resonance imaging-defined subgroups of patients with degenerative dementia. *Journal of the International Neuropsychological Society: JINS*, 15(3):459, 2009. doi: https://doi.org/10.1017/s1355617709090742.

[203] Waqas Javed and Niklas Elmqvist. Exploring the design space of composite visualization. In *2012 IEEE Pacific Visualization Symposium*, pages 1–8. IEEE, 2012. doi: https://doi.org/10.1109/pacificvis.2012.6183556.

[204] MaoQiang Xie, YingJie Xu, YaoGong Zhang, TaeHyun Hwang, and Rui Kuang. Network-based phenome-genome association prediction by bi-random walk. *PloS one*, 10(5):e0125138, 2015. doi: https://doi.org/10.1371/journal.pone.0125138.

[205] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007. doi: https://doi.org/10.1073/pnas.0701361104.

[206] Frank Emmert-Streib, Shailesh Tripathi, Ricardo de Matos Simoes, Ahmed F Hawwa, and Matthias Dehmer. The human disease network: Opportunities for classification, diagnosis, and prediction of disorders and disease genes. *Systems Biomedicine*, 1(1):20–28, 2013. doi: https://doi.org/10.4161/sysb.22816.

[207] Peter Eades and Sue Whitesides. Drawing graphs in two layers. *Theoretical Computer Science*, 131(2):361–374, 1994. doi: https://doi.org/10.1016/0304-3975(94)90179-1.

[208] Reddy Rani Vangimalla, Hyun-hwan Jeong, and Kyung-Ah Sohn. Integrative regression network for genomic association study. *BMC Medical Genomics*, 9(1):31, 2016. doi: https://doi.org/10.1186/s12920-016-0192-7.

[209] Mohammad Ghoniem, Jean-Daniel Fekete, and Philippe Castagliola. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization*, 4(2):114–135, 2005. doi: https://doi.org/10.1057/palgrave.ivs.9500092.

[210] Ronglai Shen, Adam B Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 2009. doi: https://doi.org/10.1093/bioinformatics/btp543.

APPENDIX A

APPENDIX: EXPLORATORY FACTOR ANALYSIS (EFA)

In factor analysis, each variable is expressed as a linear combination of factors, this can be represented as [117],

$$y_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \ldots + \lambda_{im} f_m + \varepsilon_i \qquad \text{(Eqn A1.1)}$$

Where $y_1, y_2, \ldots, y_p$ are variables, $f_1, f_2, \ldots, f_m$ are factors (m is always lesser than p), $\lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{im}$ are factor loadings of m factors and ε_i is a vector of error terms. In our work, we consider these *m* partitions of variables as *m* node-groupings. In matrix format, the same can be written as,

$$y = \Lambda f + \varepsilon \qquad \text{(Eqn A1.2)}$$

Where $y = (y_1, y_2, \ldots, y_p)'$, $f = (f_1, f_2, \ldots, f_m)'$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)'$, and Λ is a $[p \times m]$ factor loadings matrix. The factor analysis, primarily involves the measure of **covariances**,

$\Sigma = \mathrm{cov}(\mathbf{y}) = \mathrm{cov}(\Lambda\mathbf{f} + \varepsilon) = \mathrm{cov}(\Lambda\mathbf{f}) + \mathrm{cov}(\varepsilon) = \Lambda(\mathrm{cov}\,\mathbf{f})\Lambda' + \mathrm{cov}(\varepsilon)$, the basic assumptions of the method is $\mathrm{cov}(\mathbf{f}) = \mathbf{I}$ and $\mathrm{cov}(\varepsilon) = \Psi$ where Ψ is a $[p \times p]$ specific variance matrix that represents noise terms which is specific to each variable. Hence,

$$\Sigma = \Lambda\Lambda' + \Psi \qquad \text{(Eqn A1.3)}$$

Using spectral decomposition, a covariance matrix C can be factorized as $C = VDV'$, where V is a normalized eigenvector matrix and D is a diagonal matrix of the eigenvalues. As the matrix C is positive definite, the equation can be written as,

$$C = VD^{1/2}D^{1/2}V' = (VD^{1/2})(VD^{1/2})' \qquad \text{(Eqn A1.4)}$$

This can be equated to $\Sigma = \Lambda\Lambda'$ without an error term.

Maximum Likelihood Estimation (MLE) Method for Factor Loadings Estimation:

To execute EFA, the data should express both univariate and multivariate normal distribution [124]. The prerequisites to perform MLE method are that the data must be independent and identically distributed (i.i.d), and must have a multivariate normal distribution. For $\mathbf{x}$ being a continuous random $[p \times 1]$ vector with each variate having a normal distribution, μ is a $[p \times 1]$ mean vector of $\mathbf{x}$ and Σ is a $[p \times p]$ covariance matrix that is positive definite and symmetric, $\mathbf{x}$ can be said to have multivariate normal distribution if its joint probability density function for $\mathbf{x} \sim N(\mu, \Sigma)$, is:

$$p(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right) \qquad \text{(Eqn A1.5)}$$

The log likelihood of n observations of $\mathbf{x}$ for $L = (x_1, x_2, \ldots, x_n)$ is:

$$log(\mu,\Sigma|\mathbf{L}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| - \frac{1}{2}(\sum_{i=1}^{n}(\mathbf{x_i}-\mu)'(\Sigma)^{-1}(\mathbf{x_i}-\mu)) \qquad \text{(Eqn A1.6)}$$

$$\text{where, } \Sigma = \Lambda\Lambda' + \Psi$$

By estimating μ, Σ, and factor loadings (Λ), we can find the best fit that can maximize the log-likelihood. The equation must be solved iteratively to achieve maximum likelihood. When the diagonal matrix is $\Lambda'\Psi^{-1}\Lambda$, then we can say a unique solution is obtained, where $\Psi = diag(S - \Lambda\Lambda')$, and S is a sample covariance matrix.

For the correlation matrix R, we compute Λ using EFA and verify the fit of the model using E (residual error). The lesser the residual error, the better is the model fit the data.

$$\mathbf{E} = \mathbf{R} - \left(\Lambda\Lambda' + \Psi\right) = \mathbf{R} - \Sigma \qquad \text{(Eqn A1.7)}$$

APPENDIX B

APPENDIX: MULTI-OMICS INTEGRATIVE ALGORITHMS

Similarity Network Fusion (SNF): For the $[N \times M]$ matrix of N samples/patients and M genes, the method SNF first computes the distance matrix (W) across the samples and produces a $[N \times N]$ matrix. The number of distance/similarity matrices is equal to the count of multi-omics used for the study. The SNF algorithm iteratively updates the similarity matrix corresponding to each of the omic features as in Eqn B2.1.

$$\begin{aligned} \mathrm{P}_{t+1}^{(1)} &= \mathrm{S}^{(1)} \times \mathrm{P}_t^{(2)} \times \left(\mathrm{S}^{(1)}\right)^T \\ \mathrm{P}_{t+1}^{(2)} &= \mathrm{S}^{(2)} \times \mathrm{P}_t^{(1)} \times \left(\mathrm{S}^{(2)}\right)^T \\ \mathrm{P}^{(c)} &= \frac{\mathrm{P}_t^{(1)} + \mathrm{P}_t^{(2)}}{2} \end{aligned} \qquad \text{(Eqn B2.1)}$$

Where $P(\vec{.})$ is a normalized weighted matrix and $S(\vec{.})$ is the local affinity measured using K nearest neighbours with N_i neighbours of each node.

$$\mathbf{P}(i,j) = \begin{cases} \frac{\mathbf{W}(i,j)}{2\Sigma_{k \neq i}\mathbf{W}(i,k)}, & j \neq i \\ 1/2, & j = i \end{cases}$$

(Eqn B2.2)

$$\mathbf{S}(i,j) = \begin{cases} \frac{\mathbf{W}(i,j)}{\Sigma_{k \in N_i}\mathbf{W}(i,k)}, & j \in N_i \\ 0 & \text{otherwise} \end{cases}$$

Eqn B2.1 is repeated over t iterations to generate a P_t matrix for each omic feature. In this study, we have used two omic features hence $P_t^{(1)}$ and $P_t^{(2)}$ are generated and fused based on affinity metrics to produce a final network P^c.

To run SNF and to find hyperparameters K (number of neighbours), α (measure for local diameter), we followed the correlation measure as in (*ref:* Equation 7 [42]). We have implemented SNF with the tuned hyperparameters on both the cancer phenotypes and on the entire, and subspace data. The number of clusters in the fused affinity matrix is estimated using eigen gap and rotation cost methods.

Affinity Network Fusion (ANF): This is an improvised integrative procedure on SNF. Like SNF, ANF also computes a distance matrix for each omic feature and finds affinity measurement, but ANF works with less computation to find patient subtypes.

The K nearest neighbours Gaussian kernel, which is the combination of local Gaussian kernel and K nearest neighbours, is defined as:

$$K_{ij} = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{\delta_{ij}^2}{2\sigma_{ij}^2}} \qquad \text{(Eqn B2.3)}$$

Where δ_{ij} is the distance measure between patient i and j, and σ_{ij} is given as:

$$\sigma_{ij} = \alpha\left(\mu_i + \mu_j\right) + \beta\delta_{ij} \qquad \text{(Eqn B2.4)}$$

The Local Diameter (μ_i) of a patient i with K nearest neighbours indexes ($\mathcal{N}_k(i)$) is:

$$\mu_i = \frac{\sum_{l \in \mathcal{N}_k(i)} \delta_{il}}{k} \qquad \text{(Eqn B2.5)}$$

Normalized similarity measure between the patients, similar to Eqn B2.2, is given as:

$$S_{ij} = \frac{K_{ij}}{\sum_{j=1}^{N} K_{ij}}, \quad 1 \leq i, j \leq N \qquad \text{(Eqn B2.6)}$$

Before fusing the similarity matrices of each omic feature (S_{ij}), the matrix is further pruned, and a weighted matrix $\mathbf{W}^{(v)}$ is constructed. The final fused matrix is the summation of all pruned matrices of each omic feature. Besides, to generate a smoother fused matrix $\mathbf{W}$, the matrix is multiplied by itself for r times, where at $\mathbf{W}^{*}$, the rank of the matrix is 1, and the matrix reaches a stable state. Similar to SNF, the hyperparameters K (the number of neighbours), α (measure for local diameter), and β (measure for pair-wise distance) for ANF is estimated using the correlation measure as in (*ref:* Equation 7 [42]). The number of clusters of $\mathbf{W}^{*}$ is estimated using eigen gap and rotation cost methods.

iClusterPlus: To find subtypes of patients, we have used ‘iClusterPlus’, an improved version of *iCluster* (integrative clustering of multiple genomic data types) [210]. iCluster is a popularly known method to find the subtypes of patients. For n patients/subjects of m different omic data types with $(\mathbf{X}_1, \ldots, \mathbf{X}_m)$ omic matrices that are row normalized in the $[p \times n]$ matrix with p^m features. The subtypes are jointly estimated using latent variables of the integrative model:

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{W}_1\mathbf{Z} + \varepsilon_1 \\ \mathbf{X}_2 &= \mathbf{W}_2\mathbf{Z} + \varepsilon_2 \\ &\vdots \\ \mathbf{X}_m &= \mathbf{W}_m\mathbf{Z} + \varepsilon_m \end{aligned} \qquad \text{(Eqn B2.7)}$$

Here Z is an $[l \times n]$ matrix with l latent variables, ε is an error matrix, and $(\mathbf{W}_1, \ldots, \mathbf{W}_m)$ is coefficient matrices of m various omic features. To achieve the sparse estimation of $\mathbf{W}_m$, Lasso penalty [166] can be used, and optimal latent variables are estimated using the optimal number of clusters C. We have used the ’tune.iClusterplus’ method [188, 190] to find the optimal number of clusters (C) and the Lasso penalty (λ). The tuning method uses Bayesian information criteria (BIC) to select the best sparse model with the optimal combination of penalty parameters. The best C is computed using deviance ratio (DR), which can be interpreted as explained variation (EV). An elbow-curve between the number of clusters and % of EV reveals the optimal value of clusters of the data. Finally, patient subtypes are identified using the K-means clustering algorithm.

$$DR = (log - likelihood)\frac{fittedmodel - nullmodel}{fullmodel - nullmodel} \qquad \text{(Eqn B2.8)}$$