Advancing Graph-based Computational Approaches to Decipher Omic Signature of Diseases

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

SHREYA MISHRA

Centre For Computational Biology, Computer Science & Engineering
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

31st October 2022

THESIS CERTIFICATE

This is to certify that the thesis titled **Advancing graph based computational methods to decipher omic signature of diseases** SUBMITTED TO **IIIT-Delhi**, submitted by **Shreya Mishra**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bonafide record of the research work done by her 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.

Dr Vibhor Kumar
Thesis Supervisor
Associate Professor
Centre for Computational Biology
IIIT Delhi, 110020

Place: New Delhi

Date: 31st October 2022

ACKNOWLEDGEMENTS

An equation means nothing to me unless it expresses a thought of God.

— Srinivasa Ramanujan

This dissertation marks the end of a crucial chapter of my life, and I was fortunate to have the support of several fantastic people during its course. Each interaction taught me something unique and shaped me into the person I am today.

I would like to begin by expressing my heartfelt gratitude to my advisor, **Dr. Vibhor Kumar**, for his trust, support, and continuous boost for achieving bigger heights. It is through him that I have learned the value of persistence and self-belief; crucial qualities that will support me throughout my life and career. His compassion and kindness continues to encourage me to be a better human being, while his technical expertise and critical thinking inspire me to develop into a better researcher as well.

My sincere gratitude to **Dr. KedarNath Natarajan, Dr. Arjun Ray, Dr. Debarka Sengupta, and Dr. Angshul Majumdar**; each of whom have contributed towards the improvement of my technical writing and critical thought process via collaboration on different research papers. I would also like to thank my yearly review commiittee and comprehensive exam evaluator **Dr. Gaurav Ahuja** , **Dr GPS Raghava**, **Dr. Ganesh Bagler** for providing with their valuable feedback.

I would like to express my sincere gratitude to IIIT-Delhi, which has been like my home for the past several years. The support provided via the Institute in terms of the infrastructure, facilities, and helping staff remains at par with some of the best that I have seen till date.

I must also express my gratitude to **Dr. Pankaj Jalote** for enabling me to get the funding so that I could smoothly revive my Ph.D. program and fostering a conducive research environment at the Institute. A huge shout-out to **Priti** Ma'am for being the sole point of contact for all my administrative dilemmas and for resolving them with a

warm smile.

Finally, the past few years would have been impossible without the unconditional support of my family and friends. My parents - **Dr. Avadesh Mishra** and **Dr. Sushma Mishra**, who gave me the freedom to pursue the path I wanted to, and taught me that nothing is impossible if one has belief in themselves. You taught me to take everything with a pinch of laughter, and that has been the guiding light in this roller coaster journey.

And big thanks to my husband and fellow **Raghav Awasthi**, who has been my person throughout this journey, and continues to remain so. I could not have asked for a better life and research partner, brain-stormer, and support system. You have helped me visualize research, encouraged me whenever I felt anxious. Thanks for always believing in me. This would not have been possible without your support.

My sister, **Dr. Vartika**, brother **Harshit** and brother-in-law, **Dr. Antriksh**, who have supported me through the thick and thin, and have been my confidence across years. My niece **Anvi** and **Vedika** - their cuteness and loving words kept me going. My father-in-law **Sunil Awasthi**, mother-in-law **Greesha Awasthi**, sister-in-law **Trapti** and brother-in-law **Ankur** for supporting me throughout the PhD journey. My friends since forever - **Neha** and **Mahwish** with whom I've had several sessions about almost everything under the sun. This journey would have been incomplete without either of your support.

I am also grateful to my friends: **Smriti Chawla, Neetesh Pandey, Sarita Poonia, Omkar Chandra, Madhu Sharma, Shiju, Priyadarshini Rai, Indra Prakash Jha, Pradeep Singh, Shalini, Ridam Pal, Aditya Nagori, Samridhi, Vivek Ruhela, Chitrita Goswami, Sumit Patiyal, Nishant, Shubham, Akshaya, Anjali, Pushpita, Mitali, Ruchi Pandey**. Each one of you taught me something unique on the personal and professional front.

I am also grateful to my collaborators: **Aashi Jain, Atul Rawat and Divyanshu Srivastava** for their contribution to technical work and support needed.

This dissertation would not have been possible without you being there during the best and worst of times. Each person mentioned above has touched my life and contributed towards the fulfillment of this work. I hope this dissertation does you all proud.

Shreya Mishra

ABSTRACT

KEYWORDS: Graphs, Graph Signal Processing, Genomics, Transcriptomics, Epigenomics, Proteomics, Cancer, Diseases

Omic signatures of disease are important for personalized treatment because of the heterogeneity of diseases. Despite the advancement of computational tools, there are limited methods that can capture the latent inter-relationships between the individual components (amino acids, genes) of proteins, transcriptomic profiles. This gap may be addressed by the graph-based learning approach in a both supervised and unsupervised way which enables the creation of scientifically driven learning problems on graphs. We used graph signal processing which implements a range of tools for processing graph signal that are functions defined over the nodes in a graph. These functions represent the individual components of a biological unit. Further, these data points at the nodes are transformed into different spaces in order to bring out the latent features of the biological unit for downstream analysis. These tools elaborate on traditional signal processing and provide access to several functionalities, including filtering and frequency analysis.

In the first contribution, we devised an approach to address the noise in gene-expression profiles based on graph-wavelet driven gene-expression filtering to enhance gene-network inference. By using this approach, we were able to demonstrate how gene regulatory networks of young and elderly lung cells are different. Additionally, we contrasted differences in gene expression in lungs infected with COVID-19 with the pattern of changes in the effect of genes brought on by ageing.

In the second contribution, we have proposed a smart graph-based embedding system in our search engine (ScEpiSearch) which is capable of embedding and providing an integrative visualization of single-cell ATAC-seq profiles from various sources regardless of the species from which they originated and batch effect. Our method (scEpiSearch) calculates distance between query cells on the basis of the similarity with reference expression and epigenome cells. Here, reference cells are selected from large pool of cells based on their statistical significance of match. We demonstrated the

utility of our method in studying the lineage of cancer cells (mixed phenotype acute leukaemia) and understanding their multipotent behaviour, emphasize unique regulatory patterns in subpopulation of stem cells.

In our third contribution, we have developed a novel graph signal processing based methodology to predict biophysical properties of proteins. The model utilizes graph-wavelet of physicochemical signals of amino-acid in protein residue networks to model its biophysical properties. We demonstrate how our approach using graph wavelets can help in estimating the possible effect of disease-associated mutations on proteins using examples of prediction of globularity and folding rate.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS i

ABSTRACT iii

LIST OF TABLES ix

LIST OF FIGURES xvii

1 INTRODUCTION 1

1.1 Graph Theory Fundamentals 1

1.1.1 Centrality 2

1.2 Graph Signal Processing 4

1.2.1 Fundamentals of Graph Signal Processing 4

1.2.2 Spectral Graph Theory 5

1.2.3 Spectral Graph Wavelet Transform 6

1.3 Graph based Approaches in Improving Gene Regulatory Graphs in Transcriptomics 8

1.3.1 RNA-sequencing Fundamentals 8

1.3.2 Challenges with RNA-seq based Gene Network Inference . 10

1.3.3 Methods for Inferring Gene Regulatory Graphs 11

1.3.4 RNA-seq Denoising Methods for Better Gene-Gene Interaction Graph Inference 13

1.4 Graph Based Embedding and Multimodal Integrative Approach for Epigenomic Profiles 14

1.4.1 Challenges with Single-cell Epigenome Profiles 15

1.4.2 Graph Methods for Integrative Analysis of Multimodal Omics Data 16

1.4.3 Embedding Methods Utlizing Graph and other Techniques for Single-cell Epigenome Profiles 19

1.5 Protein Biophysical Property Modeling 21

1.5.1 Protein Graph Model 22

1.5.2 Existing Graph-based Methods for Modelling Biophysical Properties of Proteins . . . 24
1.5.3 Effect of Biophysical Properties on Mutation Sites . . . 26
1.6 Objectives and Rationale of thesis . . . 28
1.6.1 To Denoise Large Read Count Matrices of Single-cell Expression Profiles with Graph Signal Processing for Better Network Inference . . . 28
1.6.2 To Use Graph-based Integration Method for Embedding Large Single-cell Epigenome Profiles with Different Batch Effects 29
1.6.3 To Develop an Explainable Predictive Model Using Graph-Wavelet for Modeling Biophysical Properties of Proteins and Measuring Mutational Effects on Diseases . . . 30

2 Denoising Large Read Count Matrices of Single-cell Expression Profiles with Graph Signal Processing for Better Network Inference 31

2.1 Background . . . 31
2.2 Materials and Methods . . . 33
2.2.1 Spectral Graph Wavelet Transform . . . 35
2.2.2 Selecting Optimal Threshold for Graph-wavelet Coefficients 36
2.2.3 Methodologies for Inferring Gene Interaction Network . . . 37
2.2.4 Raw-Filtered Gene Network Comparison . . . 38
2.2.5 Comparison with Other Methods . . . 38
2.2.6 Differential Centrality . . . 42
2.2.7 Data Sources . . . 43
2.3 Results . . . 43
2.3.1 Assessment using DREAM Challenge Bulk Expression Profiles 44
2.3.2 Gene-networks Inference Improved by Graph-wavelet based Denoising of Single-cell Expression Profiles . . . 45
2.3.3 Age-based Regulatory Differences Shown by Improved Gene-network Inference from Single-cell Profiles. . . . 47
2.3.4 Improved Gene-network Inference Utilized to Analyze Regulatory Variations Between Young and Aged Lung Cells. . . . 49
2.4 Discussion . . . 56

3 Graph-based Integration of Large Single-cell Epigenome Profiles with Different Batch Effects 61

3.1 Background . . . 61

3.2 Material and Methods 64
3.2.1 Pre-processing Reference Profiles 64
3.2.2 Single-cell ATACseq Pre-processing to be Projected on Reference RNA-seq Profiles 65
3.2.3 Finding Matches for Query Cells in Reference Expression Profiles 66
3.2.4 Finding a Match in Mouse Reference Expression Profiles for Better Graph Learning 68
3.2.5 Graph Based Embedding of Multiple Query Single-cell ATACseq Profile 68
3.2.6 Evaluation of Co-Embedding of Single-cell ATACseq and RNAseq Profiles Across Species 72
3.2.7 Dataset Sources 74
3.3 Results 75
3.3.1 Reference Supported Improved Distance Calculation Despite Species Batch Effect 77
3.3.2 Reference Dependent Graph Embedding of Query sc-ATACseq Profiles 81
3.3.3 Case Study of Reference based Graph-embedding: Understanding Multiple Phenotype Acute Leukaemia 86
3.4 Discussion 87

4 Explainable Predictive Model using Graph-Wavelet for Modelling Biophysical Properties of Proteins and Measuring Mutational Effects on Diseases 90

4.1 Material and Methods 92
4.1.1 Weighted RIG Model 92
4.1.2 Amino Acid Features: Curation and Extraction 93
4.1.3 Spectral Graph Theory 94
4.1.4 Spectral Graph Wavelet (SGWT) 95
4.1.5 Implementation Details 96
4.1.6 Other Graph Signal Based Methods 97
4.1.7 Data Sources 99
4.2 Results 99
4.2.1 The predictive Power of Graph-Wavelet Based Feature Extraction 100

4.2.2 Comparing Graph Wavelet with Other Graph-Signal Based Methods . . . 102
4.2.3 Pattern and Importance of Graph Wavelet-based Feature Scores of Amino Acids in Determining Protein's Biophysical Property . . . 105
4.2.4 Graph Wavelet-based Features add Explainability About the Effect of Mutations on Protein's Biophysical Property . . . 107
4.3 Discussion . . . 113

5 Conclusion **118**

5.1 Summary of Contribution . . . 118
5.1.1 Denoising Large Read Count Matrices of Single-cell Expression Profiles with Graph Signal Processing for Better Network Inference . . . 118
5.1.2 Graph based Integration of Large Single-cell Epigenome Profiles with Different Batch Effects . . . 119
5.1.3 Explainable Predictive Model using Graph-Wavelet for Modeling Biophysical Properties of Proteins and Measuring Mutational Effects on Diseases . . . 120
5.2 Future Work . . . 121

LIST OF TABLES

1.1 Amino acids and their 1,3-letter codes 22

4.1 R and RMSE for modelling Protein folding rate for various methods 102

4.2 Details of Protein mutation sites for case studies 106

4.3 Benchmarking various Protein folding rate prediction methods. . . . 110

LIST OF FIGURES

1.1 By converting gene expression profiles into a gene-gene interaction matrix that represents the pairwise similarity in the expression patterns of all the genes in the system, gene co-expression networks are created. Edges in these networks could represent coincidental correlations between gene expression levels, correlations between genes that are directly regulated by the same transcription factors, correlations between genes that are regulated by the same TF, or direct regulatory relationships between genes. 11

2.1 The GWNet schematic flow. The first step is creating a KNN-based network connecting samples. For the KNN-based network of samples, a graph wavelet based filter is learnt. Graph-wavelet transform is used to filter the gene-expression of a single gene at a time. For network inference, filtered gene-expression profiles is utilized. Calculations of differential centrality between cell groups are done using the inferred network. (Mishra et. al. (247)) 44

2.2 Gene-network inference is improved by graph-wavelet based gene-expression denoising. (A) Network inference techniques' performance utilising DREAM5 challenge's bulk gene-expression data sets. Low pass filtering based on the graph Fourier transform was compared to three distinct methods of graph-wavelet coefficient thresholding. The Y-axis displays the fold change in the receiver operating characteristic curve's (ROC) area under the curve (AUC) for the overlap of the anticipated network with the ground truth set of interactions. The default value of 70% percentile was applied for the hard threshold. (Mishra et. al. (247)) . . 46

2.3 (A) A comparison of a few smoothing and imputing techniques with graph-wavelet based denoising. Fold change in the AUC-ROC for the predicted gene network's overlap with a set of known interactions For mouse embryonic stem cells, the gene-networks were predicted after the imputation or filtering of scRNA-seq profiles using various techniques. The interactions in the gold set were taken from (103). (B) A comparison of the consistency of the gene-interaction network prediction using denoising using graph wavelet with other relevant smoothing and imputing techniques. Here, the Phi (ϕ_s) score was used to forecast the gene regulatory network. To test resilience against the batch effect, inferred networks from two scRNA-seq profiles of mESC were compared. (C, D) Reliability in the gene-interaction network prediction for mESC using two batches of scRNA-seq data. Results for Pearson correlation and a co-expression network based on ρ scores are shown below. (Mishra et. al. (247)) 48

2.4 Performance and noise analysis for pancreatic single-cell RNA-seq profile.(A) Performance as measured by the estimated network's overlap with the data set on protein-protein interactions. (B) A consistency assessment of the anticipated network. It's crucial to minimise noise-related discrepancies when comparing two networks. The results for the correlation-based co-expression network are shown below. (C) This figure illustrates the differences in gene expression across single cells both before and after denoising (filtering). Gene variation in a cell type was determined independently for each of the three phases of ageing (young, adult and old). Compared to young alpha and beta cells, the variance (estimated noise) is larger in older alpha and beta cells. However, after denoising, the gene variance in each stage of ageing becomes equal. (D) Here, the impact of noise on the calculated differential centrality is seen. Here is a comparison of the estimated gene network degrees for elderly and young pancreatic beta cells. Denoised expression-based networks predict fewer non-zero differential degrees than unfiltered expression-based networks do. (E) Enriched panther pathway terms for the top 500 genes in old and young pancreatic beta cells that exhibited the significant reduction in variance after denoising. (Mishra et. al. (247)) 50

2.5 Improved regulatory inferences from single-cell transcriptome pre-processing of ageing lung cells using graph-wavelet (A) Reliability of networks predicted using the scRNA-seq profiles of young and elderly lung cells from Kimmel et al. . This figure show the coverage of the top 10,000 edges in young cells in the network inferred from old cells. After graph-wavelet based filtering, the estimated networks for old and young cells with the same type seem to have more overlap. The term "Raw" here denotes that the unfiltered scRNA-seq profiles were used to infer both networks (for old and young). Whereas the same result from the filtered scRNA-seq profile is shown. Utilizing correlation-based co-expression, networks were inferred. (B) The network overlap inferred from two distinct data sets, each with their unique batch effect, is shown. The X-axis displays the number of edges in the network inferred using the data set from Angelidis et al (GEO Id: GSE124872). The percentage of the top 10,000 edges in the network calculated using the Kimmel et al. data set is shown on the Y-axis. (C) The top 500 genes with the 10 most enriched Panther pathways have greater PageRank in young AT2 cells than in elderly AT2 cells. (D) The top 1000 genes with the 10 most enriched panther pathways had greater PageRank in older AT2 cells than in younger ones. (E) Scatter plot showing the difference in transcription factor (TF) PageRank (old-young) computed using networks predicted for old and young AT2 cells from the Kimmel et al. data set. Only TFs having a differential degree that is not zero are shown. (Mishra et. al. (247)) 54

2.6 Panther pathway keywords were enriched for genes with substantially increased differential expression (FDR < 0.1) in aged AT2 cells vs young cells. The bar in grey indicates negligible enrichment (Pvalue < 0.05). Differential expression analysis results do not include some of the phrases that appear in PageRank-based results. RAS Pathway, JAK/STAT Signaling, and Cytoskeletal Regulation by Rho GTPase are examples of phrases that did not seem to be enriched for genes with greater expression in elderly AT2 cells. (Mishra et. al. (247)) 56

2.7 In both old and young mouse At2 cells, Jun genes interact with other genes. Ribosomal proteins make up the majority of Jun's interaction partners. Additionally, *Etv5* and *Jund* seem to be dependent on c-JUN via co-expression. (Mishra et. al. (247)) 57

2.8 Analysis of scRNAseq profile in SARS-COV-2-infected lungs (COVID). (A) PageRank distribution of genes up-regulated in COVID-infected lung ($FDR < 0.05$) (122). For networks that were computed using scRNA-seq of both young and elderly AT2 cells, the PageRank is shown. For genes that were downregulated ($FDR0.05$) in COVID-infected lung, PageRank is shown in a similar fashion.(B) Top 10 Panther pathway genes that are abundant in COVID-infected lung. In older AT2 cells, the phrases indicated by an asterisk (*) likewise significantly enrich for higher pagerRank genes. (C) Top 10 wiki pathway for genes that have higher expression in COVID-infected lungs. In older AT2 cells, the phrases indicated by an asterisk (*) are likewise enriched ($Pvalue0.05$) for genes with higher pageRank. (D) The top 3 known transcription factors (TF) motif enrichments in gene promoters in COVID-infected lung. (E) Fold change in expression of genes with positive and negative correlations to transcription factors in aged AT2 cells in lung with COVID infection. Two transcription factors (TFs) *Etv5* and *Stat4*, which have increased PageRank in elderly AT2 cells, are shown. The results for *Erg*, which had higher PageRank in young AT2 cells, are also shown as a control. The majority of the genes whose expression correlated positively with those of *Etv5* and *Stat4* in aged AT2 cells were also up-regulated in COVID-infected lung. *Erg*, meanwhile, perceives the reverse trend. In aged AT2, genes with positive association to *Erg* genes were more down-regulated than those with negative correlation. Such findings imply a potential role for transcription factors (TFs) whose impact (PageRank) increases with age in either activating or posing the genes up-regulated in COVID infection. (Mishra et. al. (247)) 58

3.1 An illustration of the proposed method in scEpiSearch for improved regulatory state inference and annotation of query scATACseq utilising a large pool of single-cell epigenome data sets. It consists of the following steps: Expression Reference Data Preparation, Query Processing, Mapping and Projection Based Cross-Species and Cross-batch Query embedding. The cross-species and cross-batch query embedding provides co-embedding of various open-chromatin profiles utilising existing reference single-cell profiles, regardless of variation in peak-list in read-count matrix, batch effect, and species. (Mishra et. al. (248)) 75

3.2 A) Evaluation of scEpisearch in comparison to 3 other methods based on the retrieved gene scores correlation for matching query scATACseq to a collection of reference single-cell scRNAseq. 10,000 MCA (mouse cell atlas) cells were selected as the scRNAseq reference dataset (MCA). Here, accuracy displays the proportion of query cells with the correct cell-type among the first 5 matches.B) Accuracy using the scEpiSearch for queries on the scATACseq read-count matrices is shown as follows, from left to right: i) query human scATACseq to reference human single-cell expression ii) search the mouse reference scRNAseq using the mouse scATACseq. iii) cross-species search – reference mouse scRNAseq using human scATACseq. The proportion of query cells for which the right annotation was one of the top 5 results is shown on the Y-axis.For the scEpiSearch modules for both faster and accurate, accuracy is shown as bar-plots. (Mishra et. al. (248)) 76

3.3 (A) A comparison of scEpiSearch with integrative approaches, using scRNA-seq profiles of 10,000 cells from the mouse cell atlas. Here, the search was restricted to the scATACseq profiles of three mouse cell types: B cells, macrophages, and endothelial cells. T he silhuotte index of the query cells for vicinity to the correct reference cell-types is shown. (B) Analysis of cross-species integrative approaches and scEpiSearch employing reference scRNAseq from MCA and human PBMC scATACseq profiles as the query. Also, human PBMC silhouette coefficients are shown. For the purposes of calculating the silhouette coefficients, immune cells and query cells were assumed to belong to single class in the references, whereas other cell types were regarded to be in other class. (Mishra et. al. (248)) 77

3.4 Utilizing reference expression and the query scATACseq from mouse cells, scEpiSearch is compared to integrative approaches. A) Mouse endothelial scATACseq profile was the query (Cusanovich et al.). For each of the 4 approaches, the query cells' silhouette coefficients are shown. B) Mouse macrophage epigenome profiles query (Cusanovich et al.) are used. Also shown are the silhouette coefficients for the query cells. C) Incorrect assessment of the 2D embedding figures when all cells, including the reference MCA cells silhouette coefficients are considered. Mouse endothelial cells, mouse macrophages, and 3-cell combinations are shown by the corresponding labels on the subfigures. (Mishra et. al. (248)) 78

3.5 Utilizing the reference scRNAseq of mouse cells as the basis for comparison and the scATACseq profile of human cells as the query A) The search comprised of human embryonic stem cell single-cell ATAC-seq profile (H1ESC). Additionally provided are the query cell silhouette coefficients for the four approaches. B) The scATACseq of human neuronal cells were query. Also included are the silhouette coefficients for the query cells. C) The silhouette coefficients presented for all cells in 2D embedding plot, including the reference MCA cells. According to the labels on the corresponding sub-figures, H1ESC, human neurons, or human PBMC were the source of the single-cell ATAC-seq profiles. (Mishra et. al. (248)) 79

3.6 Evaluation of embedding of query sets of single-cell open-chromatin profiles irrespective of batch effect, species, differences in peak-list and their source. Embedding plot from ScEpiSearch derived from projections onto mouse expression profiles. (A) Queries consisted of scATACseq profiles of human-neuron, mouse-neuron, human-HSC, house-HSC, human-Myoblast, human-GM12878 cells from two batches and mouse-B-cells. The purity of density based spatial clustering (using DBSCAN) with embedded coordinates is shown here in terms of ARI (Adjusted Rand Index) and NMI (normalized mutual information) scores. (B) Queries made for Human-GM12878 cell, Mouse B-cell, Human-HEK293T, Mouse-Proximal tubule. (C) Queries were made for Human-GM12878, Mouse B-cell, Human T-cell, Mouse T-cell. (Mishra et. al. (248)) 82

3.7 Evaluation of query scATACseq profiles from various species and batches using 2D embedding. a) Query for Human-Neuron, Mouse-Neuron, Human-HSC, and Mouse-HSC are made. None of the other techniques examined in this study were able to provide accurate low-dimensional embedding like scEpiSearch. For others,TSNE plot is shown. After utilising labels HSC and forebrain/neurons and applying DSCAN to the 2D coordinates, the right-bottom panel displays clustering-purity in terms of ARI and NMI. b) This plot displays the evaluation of 2D embedding plots showing the computation of silhouette coefficients. (Mishra et. al. (248)) 83

3.8 Using scEpiSearch's 2D embedding to follow the dedifferentiated state of blood-cancer patients' leukaemia cells. A) Pie chart of top scRNAseq types matches for scATACseq profile of blood cells from individuals with multiple phenotype acute leukaemia (MPAL) is shown in Figure 1. (GEO id: GSE139369). B) Pie-chart displaying cell types of human peripheral blood mononuclear cells (PBMCs) with top matching scRNAseq cells (GEO id: GSE139369) C) scEpiSearch based 2D embedding of scATACseq of 3 kinds of cells: peripheral blood mononuclear cells (PBMC) from healthy cells, progenitors of blood cells, and blood cells taken from patients with mixed phenotype acute leukaemia (MPAL). Granja et al. released the scATACseq profiles of MPAL and PBMC cells, and progenitor cell scATACseq are from a different study. Most PBMC cells are located further from progenitor cells and closer to B cells in the embedding plot created by scEpiSearch. Progenitor cells are further distant from MPAL cells. Some MPAL cells are so thoroughly dedifferentiated that they even entirely overlap with blood cell progenitors. the identical 2D embedding result, but the MPAL cells were coloured in accordance with the origins of the patient (P1 and P2). MPAL cells from two patients have some overlap, but they also have various degrees of dedifferentiation. D) Outcomes from other tools for 2D embedding of scATACseq of three different kinds of cells: peripheral blood mononuclear cells (PBMC) from healthy persons, progenitors of blood cells, and blood cells from patients with mixed phenotype acute leukaemia (MPAL). Other approaches either failed to colocalize B and T cells with PBMC or jumbled up the location of several hematopoietic progenitor cell types. (Mishra et. al. (248)) 84

4.1 The pipeline for ProteinGW is shown in a flowchart. For the modelling of each protein function, protein structures were acquired from PDB database. For each protein, a dictionary was made that included the distance matrix, sequence, etc. Adjacency matrix known as weighted RIG was constructed. Then, using a weighted-RIG matrix, a network tailored to each protein is constructed. Additionally, each amino acid (node) is represented on the graph as a signal with its physical, chemical, and network characteristics, and the ideal threshold is determined to cutoff wavelet coefficients. An ML model is then trained. (Mishra et. al. (249)) 98

4.2 Performance of several machine learning models after feature extraction using the graph-wavelet method utilising five fold cross-validation. For the purpose of modelling the categorization of transmembrane and globular proteins, (A) MCC, Accuracy, Macro-F1, and ROC-AUC are compared. XgBoost, AdaBoost, KNN, Gaussian Naive Bayes, logistic regression, SVM, and random forest are machine learning models that are compared.For modelling the classification of soluble and insoluble proteins by machine learning models, (B) MCC, Accuracy, Macro-F1, and ROC-AUC are compared. (C) MCC, Accuracy, Macro-F1, and ROC-AUC are compared for modelling classification of all-alpha and all-beta proteins. (Mishra et. al. (249)) 103

4.3 A) correlation value (R) for estimating the rate of protein folding is presented. The machine learning models are supplied with features taken from ProteinGW. Comparisons are made between ElasticNet, Decision Tree, Random Forest, KNN, SVR, Ridge, Lasso, and Linear Regression. (B) The plot shows models' Root mean squered error (RMSE) FOR protein folding rate prediction. (Mishra et. al. (249)) 104

4.4 Comparison of ProteinGW with other methods and feature-weights for their amino-acid properties. (A) Accuracy of protienGW, convolutional neural network (CNN) and graph Fourier Transform (GFT) at the different number of training data points, The training fraction is reduced from 0.85 to 0.50. The importance of the feature score was calculated using a Random forest-based model for Solubility,Transmembbrane-Globular, All alpha-all beta (B) Similarly, AUROC is shown. (C) Here, F1-score is shown for all 3 properties. (Mishra et. al. (249)) 108

4.5 Benchmarking of Protein Solubility prediction methods. 109

4.6 Graph wavelet-based features scores (GWFS) on various scales in simulating protein biophysical properties A) Protein-folding rate modelling with four wavelet scales scores is illustrated. Plot shows that at scale 4 (corresponding to low frequency), conservation score, node weighted degree, refractive index, node degree, and residue count are critical, but polarity is more significant at scale 1. Similar to this, polarity at scale 3 also seemed to be important. (B) 4 wavelet scales for distinguishing all-alpha and all-beta proteins are shown. The most relevant features were molecular weight, coil propensity, compressibility, and bulkiness at scale 4. In a similar way, (C) polarity, molecular weight at scales 1 and 4, conservation score at scale 3, coil tendency at scale 4, refractive index at scales 1 and 2, and turn tendency at scales 1 and 3 have shown to be more significant for transmembrane-globular. (D) Similarly, the refractive index at scale 4 and flexibility, node degree, residue frequency, and partial specific volume at scale 1 seemed to be more significant for the categorization of soluble and insoluble proteins. (Mishra et. al. (249)) 112

4.7 Transmembrane/globular property in effect of disease causing mutation(A)Percentile Scores of top features for proteins PDB:3DZQ(EPHA3),2BBA(EPHB4),6Y
gene).Average value for features across all four wavelet spectrum is shown.(B)Probability of proteins before,after mutation of original class.For control,average change in probability for mutations at 10 random sites is shown for every protein.(C)Visualization of 4Y63 mutation site. (Mishra et. al. (249)) 114

4.8 Assessing impact of disease-causing mutation on folding rate(A) Percentile Scores for top features for proteins 2A5E (Met53Ile, Melanoma), 1V9E (His94Tyr, Osteopetrosis), 2A5E (Val95Ala, Non-small cell lung carcinoma), and 2A5E(Glu119Gln, A biliary tract tumor) are bulkiness, conservation score, flexibility, refractive index, clustering coefficient, partial specific volume, residue frequency, and molarity . (B) The difference in rate of protein folding before and after a mutation. Every protein's average change in likelihood due to mutations at 10 random locations is also shown. (C) A visual representation of the 2A5E mutation site (Met53Ile, Melanoma, and Cutaneous Malignant) and the anticipated, actual, and changed protein folding rates. (Mishra et. al. (249)) 115

CHAPTER 1

INTRODUCTION

1.1 Graph Theory Fundamentals

Network topology and properties are a popular topic in discrete mathematics and computer science (2). Graphs may be utilized to depict a wide spectrum of physical, social, and biological framework. They are commonly used to express pairwise relationships among a set of components. Graph theory is concerned with the statistical and computative tools and methodologies that can be utilized to study and analyse graphs.

Formally, Graphs can be defined as $G(V, E, W)$ constituting finite compilation of vertices (V) (also known as nodes) linked by an ensemble of edges E. Vertices of the network/graph are usually labeled as $v_1, v_2, ..., v_N$ with $|V| = N$. The set of edges E connecting these vertices, such that $e_{i,j}$ would be an edge associating vertex v_i to vertex v_j.

The weighted adjacency matrix W, which defines the strength of connections among the vertices. If there exist an edge $e_{i,j}$ in graph, $w_{i,j}$ represents its weight. For two vertices v_i and v_j not directly connected, $w_{i,j} = 0$.

Graphs can be categorised in a variety of ways. A graph can be characterised as undirected or directed on the basis of direction of its edges. Connections among vertices(nodes) in an undirected graph are oriented from one vertex to the next. Graphs can be categorized as weighted or unweighted based on the weights assigned to the graph's edges. An unweighted graph has all of its edges with the same weight. $W_{i,j}$ for such a graph is either 1 if v_i is linked to v_j, or 0 otherwise. A graph can be connected or disconnected based on its connectivity. If there exist a route connecting each pair of vertices in the network, it is believed to be connected; otherwise, it is said to be unconnected.

A vertex's degree is defined as number of nodes connected directly to the given vertex. Let deg_i indicate vertex v_i's degree. A vertex's weighted degree would be the total of weights of every edge adjacent to it.

The vertex's clustering coefficient is defined as the fraction of number of associations within its neighbours to the aggregate number of possible associations within its neighbours. Clustering coefficient of a vertex in a simple graph is computed as

$$C_i = |e_i, j| \tag{1.1}$$

1.1.1 Centrality

Due to the extremely heterogeneous structure of complex networks, certain nodes may be regarded more important than others. However centrality can also be defined in respect of payload that a particular vertex endures. The definition of centrality is not universal and varies depending on the context. Several metrics of centrality have been developed, each of which takes into account a different notion (4).

The number of connections connected to each node determines its degree centrality. Since a densely linked node (hub) may not be central, the degree centrality might be regarded a local centrality metric. With respect to adjacency matrix (A), the total of components in row i of matrix A may be used to compute degree based centrality of the vertex i i.e.

$$k_i = \sum_{j=1}^{N} A_{i,j} \tag{1.2}$$

Here, number of vertices in graph are denoted by N.

Closeness centrality (3) can be defined as a total of edges in the shortest path connecting vertices i , j determining the distance between them. In terms of distance, a central node would be adjacent to every other vertex/node in the graph. The closeness centrality for i (average length of i with other nodes) is defined mathematically as

$$C_i = \frac{N}{\sum_{j=1, j\neq i}^{N} d_{i,j}} \tag{1.3}$$

where $d_{i,j}$ would be the shortest distance between i, j, and N would be a number of vertices in graph.

Centrality can also be defined in terms of load, if we examine the movement of particles on a network (betweeness centrality) (3). The total number of shortest pathways

traveling through a node determines the load in node i. However, there might be more than one smallest length between vertices a, b, load in vertex i is better defined being the proportion of smallest pathways linking every set of vertices $(a, b)_{a,b=1,...,N}$. Hence betweeness centrality of node i,

$$B_i = \sum_{(a,b)} \frac{\eta(a, i, b)}{\eta(a, b)} \tag{1.4}$$

Here $\eta(a, i, b)$ is total of smallest paths between nodes a , b crossing along nodes i. $\eta(a, b)$ would be a total of smallest length inbetween a , b. The total includes all combinations of unique vertices (a, b). In this instance, several pathways pass a centralized node, which produces the largest value of B.

PageRank (5) is a modification of Eigen Vector centrality and can be applied to any kind of network. Three discrete factors that determine PageRank are 1. The number of incoming links 2. Link propensity of linkers 3. The centrality of linkers. The PageRank of a non-directed graph can be defined as: Let R be the PageRank vector and D would be distribution array, then

$$D = \frac{1}{2|E|} \left[deg(p_1)\ deg(p_2)\ .....\ deg(p_N)\right] \tag{1.5}$$

Here $deg(p_1)\ deg(p_2)\ .....\ deg(p_N)$ would be vertices' degree. E would be set of edges of graph. N is the number of pages. Hence, as proved by Grolmusz (5), $v = \frac{1}{N}.1$ gives,

$$\frac{1-d}{1+d}||v - D||_1 \leq ||R - D||_1 \leq ||v - D||_1 \tag{1.6}$$

Here, d is the damping factor and D is degree distribution array.

Thus, undirected graphs' PageRank is equal to degree distribution array if (only if) the network is regular, implying every vertex has the same degree.

1.2 Graph Signal Processing

1.2.1 Fundamentals of Graph Signal Processing

DSP (discrete signal processing) has shown to be very effective in operating/analyzing time signals including speech, communications, econometric time series, radar, time-space signals, and space-dependent signals (such as images as well as multidimensional signals as seismic, hyperspectral information). Data referenced by a network's vertices is referred to as a signal on a graph or just a signal. This technique is referred to as DSP on graphs (1).

Graph Signals

Considering dataset with size n with a few relational parameters about the identified data components. A network $G = (V, A)$ could be employed to depict this data, wherein $V = v_0, ..., v_N$ would be set of vertices, A would be graph's weighted adjacency matrix. Every dataset constituent a vertex v_n, and the weighted $A_{n,m}$ corresponds to directed edge between v_m and v_n represents relationship degree between n^{th} and m^{th} component. G is a directed, weighted graph in general, because data elements can be connected to each other in a multitude of ways. The edge weights $A_{n,m}$ don't have to be non negative reals, they can be any real or complex number (data elements could be negatively correlated). The neighbourhood of v_n is defined by $N_n = m|A_{n,m} \neq 0$ and consists of the indices of nodes linked to v_n. Hence, graph signals can be defined as vector $s = (s_0, s_1, ..., s_{N-1})^T$.

Filters on Graphs

Filters are approaches which accept input in the form of a signal and create an other output signal — are used in traditional DSP to process signals. In DSPG, a similar idea of graph filters exists for graph signals where Linear,shift-invariant filters, an extension of linear time-invariant filters applied in time series Digital signal processing, are considered.

Similar to traditional DSP, applying filters on a network is accomplished via matrix

and vector multiplication. A graph filter $H \in C^{NxN}$ which produces output Hs for input $s \in S$ representing linear method. Given that the linear combination of the filter's outcomes for each signal's input and output signals is linear,

$$H(\alpha s_1 + \beta s_2) = \alpha H s_1 + \beta H s_2 \tag{1.7}$$

Additionally, shift-invariant graph filters apply graph shift to output corresponding to employing graph shift to input before filtering such that

$$A(Hs) = H(As) \tag{1.8}$$

1.2.2 Spectral Graph Theory

In spectral graph theory, methodologies and concepts from traditional graph theory are extended to matrices' eigen values, characteristic polynomial and eigen vectors (for e.g., the adjacency matrix) relating to graphs.

Weighted Graphs

The components of a weighted graph are a collection of nodes/vertices (V), a pool of edges (E), a weighted function $w : E \implies R^+$ that gives each edge a non-negative weight. Here, considering only the finite graphs. A NXN 2-D array with entries $a_{m,n}$ serves in form of adjacency matrix (A) considering a weighted graph G.

$$a_{m,n} = \begin{cases} w(e), & \text{if } e \in E \; connects \; vertices \; m \; and \; n \\ 0, & \text{otherwise} \end{cases} \tag{1.9}$$

The degree pertaining to every node m in a weighted graph, denoted by $d(m)$, which is equal to total of all edge weights that intersect the given node. It suggests that $d(m) = \sum_n a_{m,n}$. It can also be specified that matrix D contains zeros everywhere else and diagonal elements equal to the degrees.

Just like $L = D - A$, a non-normalized Laplacian, an alternate normalized version

of the Laplacian given as (218)

$$L^{norm} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}AD^{-1/2} \tag{1.10}$$

It can also be emphasized that matrices L and L_{norm} can not be compared, particularly in terms of the eigen vectors.

Graph Fourier Transform

In graph Fourier transform (218), graph Laplacian L has all possible orthonormal eigenvectors given it is real symmetric matrix. For $l = 0, ..., N-1$, we designate them as χ_l, and the corresponding eigenvalues are λ_l.

$$L_{\chi l} = \lambda_l \chi_l \tag{1.11}$$

Since each of the λ_l is real, L is symmetric. It has been shown that for the graph laplacian, all positive eigenvalues and given that zero occurs in terms of eigen value along with frequency equals a total number of connected graph components (132). Assuming graph G is connected, eigenvalues can be arranged such as

$$0 = \lambda_0 < \lambda_1 \leq \lambda_2 ... \leq \lambda_{N-1} \tag{1.12}$$

For a given function $f \in R^N$ on nodes of G, Graph Fourier transform is defined as

$$\hat{f(l)} = \langle \chi_l, f \rangle = \sum n = 1^N \chi_l^*(n) f(n) \tag{1.13}$$

Inverse Fourier transform is defined as

$$f(n) = \sum_{l=0}^{N-1} \hat{f(l)} \chi_l(n) \tag{1.14}$$

1.2.3 Spectral Graph Wavelet Transform

The kernel function, which would be comparable to the Fourier domain wavelet $\hat{\psi}^*$, will decide the wavelet transform: $g : R^+ \implies R^+$. The kernel g would function in form of band-pass filter since $lim_{x \implies \infty} g(x) = 0$ and it fulfils $g(0) = 0$.

Wavelet

As operator-valued expressions of Laplacian, wavelet operators provide spectral graph wavelet transform. The continuous functional calculus may be used to create measurable function of bounded self-adjoint linear operator on Hilbert space (133). The Spectral representation of operator, which in this context is equal to graph Fourier transform described in the preceding section, is used to achieve this. The wavelet operator $T_g = g(L)$ specifically modulates each Fourier mode as it works on a given function f for the spectral graph wavelet kernel g.

$$T_g\hat{f}(l) = g(\lambda_l)\hat{f}(l) \tag{1.15}$$

Implementing inverse transform gives

$$(T_g f)(m) = \sum_{l=0}^{N-1} g(\lambda_l)\hat{f}(l)\chi_l(m) \tag{1.16}$$

The equation $T_g^t = g(tL)$ defines the wavelet operators at given scale t. Afterwards, localising such operators and implementing these on impulse on a particular node, the spectral graph wavelets are produced as,

$$\psi_{t,n} = T_g^t\theta_n \tag{1.17}$$

Explicitly extending this in field of graph demonstrates,

$$\psi_{t,n} = \sum_{l=0}^{N-1} g(t\lambda_l)\chi_l^*(n)\chi_l(m) \tag{1.18}$$

Thresholding methods

- Hard Thresholding: This is a linear function that, according to (134), discards coefficients that are less than cutoff score which is based on variance of noise. The hard thresholding based filter parameters can be provided by (135),

$$h_h(i) = \begin{cases} 1, & \text{if } |y(i)| > \tau \\ 0, & \text{otherwise} \end{cases} \tag{1.19}$$

  It generally follows the "keep or kill" principle. When the noise terms do not surpass the threshold in this technique, false "blips" and discontinuities show up in the output.

- Soft Thresholding: This non-linear function would be comparable with a hard threshold however, it decreases larger valued wavelet coefficients over threshold (134). Relative to the hard threshold, this results in a more continuous and smooth output. The soft threshold based filter coefficients can be taken (135) as i.e,

$$\eta_t(y) = sgn(y)[|y| - t] \tag{1.20}$$

The universal threshold value is taken into account for both hard and soft thresholds, would be provided through $t_n = \gamma_1 \sigma \sqrt{2log(n)/n}$, where n is input signals' length, $\gamma_1 isaconstantand\sigma$ would be an estimate of variance made using coefficients at the highest degree of detail.

- BayesShrink: BayesShrink (137) (138) employs Bayesian computational framework which can determine optimum cutoff which minimises Bayesian associated risk, which is defined for wavelet coefficients in form of the generalised Gaussian distribution in every specifics of the sub band.

$$\sigma_B = \frac{\lambda^2_{noise}}{\lambda_{signal}} = \frac{\lambda^2_{noise}}{\sqrt{max(\lambda^2_G - \lambda^2_{noise}, 0)}} \tag{1.21}$$

Here, $\lambda^2_G = \frac{1}{P_s} \sum_{x,y=1}^{Ns} V^2_{xy}$, P_s would be number of wavelet coefficients V_{xy} on sub band being considered.

- SureShrink: To choose the ideal threshold value, Dohono and Johnstone (136) suggested using the sure shrink approach (134). The Stein's unbiased estimator of risk serves as the foundation for this approach. This approach provides an objective estimate of the loss $||(\hat{\mu} - \mu)^2||$. The SURE (quadratic loss) function is used to evaluate the risk for a certain threshold tau. At this point, cutoff score can be chosen by reducing risks associated with tau. For large samples of data, this technique has improved MSE characteristics. The unbiased risk assessment provided by the stein can be defined as

$$R_s(\tau) = N + E\{||g(y)^2|| + 2\nabla.g(y)\} \tag{1.22}$$

This threshold value can now be chosen taking into account set $y_0, y_1, ..., y_{N-1}$ minimal's risk value.

$$surethreshold(\tau) = arg\, min R_s(\tau)\, where\, \tau = y_0, y_1, ...y_{N-1} \tag{1.23}$$

1.3 Graph based Approaches in Improving Gene Regulatory Graphs in Transcriptomics

1.3.1 RNA-sequencing Fundamentals

Much of the progress in understanding disease phenotypes has come from analyzing gene transcriptional data. It is achieved by taking static indicators of the abundance of

RNA levels in different cellular states and using these measurements to develop network models showing the transient processes driving biological systems. However, it is argued that transcriptomic data do not adequately depict the complexities of gene regulation or the kinetics of biological activities. But, given the limitations of the experimental limitations, gene expression data are the finest resource available today for modeling regulatory networks and developing predictive models of cellular response in a variety of conditions.

Transcriptome can be defined as collection of entire RNA molecules contained within cells, notably messenger RNA (mRNA). The abundance levels of these molecules referred to popularly as gene expression are often employed as the key input variables for algorithms that attempt to predict transcriptional networks.

Despite the fact that inference of cellular networks is one of the areas where most substantial advancements have been made in terms of model development. These models may be useful in the field of network medicine, still it remains both a problem and an active research topic (6) (99). This research has highlighted the need of having integrated datasets that capture the numerous components contributing to the gene regulation process. Fortunately, new DNA-sequencing methods are enabling the collection of increasingly massive and complicated datasets from individual samples, including genome-sequence data, transcriptome data, and genome-wide data on epigenetic modification patterns.

RNA-seq begins by identifying the base sequences of individual RNA molecules, which are then projected to reference genome to calculate abundance of specific gene transcripts. RNA-seq data, in contrast to continuous values recorded with microarrays, reflect counts of RNA sequence reads that relate to a specific gene. One benefit of this technique is that it can, in theory, monitor the stages of expression of each and every gene. However, many variables, including gene sequence, biases in sequencing library creation, the total quantity of sequenced reads, and the technique used to map sequence reads to genes, can all have an influence on the efficacy and precision of RNA-seq measurements.

The complexity of biological systems contributes to the difficulty of correlating genetic information to observable phenotypes. With reference to gene regulatory networks, transcription reveals an underlying process in which the concentration of mRNA

in a particular cell or pool of many cells is regulated by context-specific activity of a range of regulating factors. Transcriptional activation of a specific gene begins with the collective binding of various transcription factors in special control areas of the DNA which are mostly located immediately before the target gene, in its promoter, but they can also be located distally, in what are known as enhancers.

1.3.2 Challenges with RNA-seq based Gene Network Inference

Gene regulation cannot be viewed as a single interaction or even a single route, but instead as a complex network of interconnected genes and gene products. One approach to think about these regulatory interactions is network, where genes and TFs are denoted as nodes associated with one other by directed edges. In general, this network and the interactions between its components are to be understood since this is what underpins how organisms respond to environmental and other signals and may dictate how functioning biological systems are changed when disorder occurs. Understanding and deducing these networks properties has become one of many intriguing and demanding area of research in computational and systems biology.

Even in a homogenous cell population, scRNAseq profiles frequently exhibit variation in levels. It is possible to identify dominant pathways in a cell type and infer regulatory networks between genes using this heterogeneity. Due to the sparseness and ambiguity in the pattern of gene expression in single-cell RNA-seq profiles, however, the optimum metrics for gene–gene interactions remain uncertain. Regulatory mechanisms involved in diseases and development is being widely understood through gene-regulatory network inference and are used for system-level modeling. The edge weights among nodes (edge weights indicating regulatory interactions between genes) represent interdependences between the network's variables. Gene-Gene interaction networks may be beneficial in inferring causal models (83) (84), comprehend perturbation studies, comparative analysis (85) (11), and identify new drugs (87) (13). Many approaches for estimating node interdependence have been developed with majority of approaches relying on mutual information (MI), pairwise correlation or other similarity metrics across gene expression data from distinct conditions or time points. However, with patterns exhibiting low but effective background similarity, the resultant edges are frequently impacted by indirect dependencies. Even when there is a meaningful inter-

action between two nodes, their influence and strength are frequently underestimated owing to noise, background-pattern similarity, and other indirect interactions.

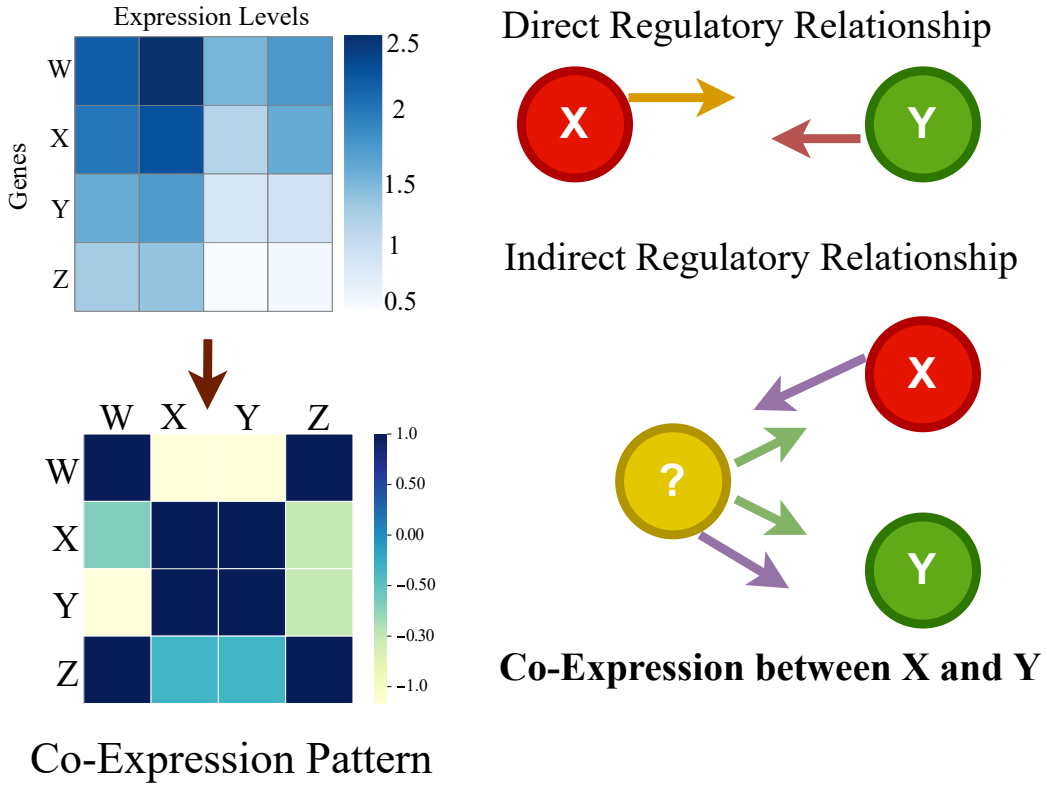

Figure 1.1: By converting gene expression profiles into a gene-gene interaction matrix that represents the pairwise similarity in the expression patterns of all the genes in the system, gene co-expression networks are created. Edges in these networks could represent coincidental correlations between gene expression levels, correlations between genes that are directly regulated by the same transcription factors, correlations between genes that are regulated by the same TF, or direct regulatory relationships between genes.

1.3.3 Methods for Inferring Gene Regulatory Graphs

The Methods for inferring co-expression networks (14) (15) generally begin by calculating a similarity score, such as a Pearson correlation coefficient, between each pair of genes by comparing their expression levels across samples. Then, a threshold value is specified for the lowest value at which genes are considered correlated. If we set every element in the matrix to either zero (below the threshold) or one (at or above the threshold), we get an adjacency matrix with rows and columns representing genes and matrix entries representing the existence or absence of an edge linking them. Sknnider et al. (13) used seventeen measures of association to build a network of gene interactions. In their research, they found that utilising scRNA-seq profiles, two metrics

of association—phi and rho—performed the finest at predicting gene-gene interaction (co-expression based).

The correlation-based similarity matrices are symmetric across the diagonal, the networks created by these metrics are often undirected. They also incorporate information about every pair of genes rather than simply TF-target gene interactions; hence, regulatory linkages are mixed along with coregulatory correlations (Figure 1.1). Edges connecting two genes, neither of which is a TF, can be trimmed to generate a putative regulatory network from this matrix. To remove implausible associations, more specificity may be applied to the remaining edges using TF binding site motifs, and directionality can be assigned by assuming that edges point from TFs to non-TFs. It quickly became clear that networks built in this manner did not correctly represent the underlying regulatory mechanisms.

WGCNA (89) was developed for solving the latter problem. In WGCNA, the calculated co-expression levels between pairs of genes are adjusted by employing a power adjacency function and taking the absolute value of the correlation to a power i.e,

$$a_{ij} = |cor(x_i, x_j)|^{\beta} \tag{1.24}$$

The correlation values are limited to a maximum of one, the least correlated gene pairs will have their weight converge to zero as the value of β increases, while perfect correlation will remain unaffected. It is crucial to highlight that WGCNA is especially designed to find set of co-expressed genes with improved precision in the regulatory network.

Several scientists realized that biological relationships may be nonlinear and that simple linear measurements would overlook them. Alternative approaches for capturing such complicated interactions and inferring networks have been developed, which employ correlation measures such as mutual information (MI) (16) (18) (19). However these measures can capture more complex information, they often require large datasets to approximate the right underlying probability distribution and infer connections effectively.

A method ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) (87) was introduced that used MI to reconstruct a gene regulatory network.

ARACNE considers that a high mutual information value between two genes can be the consequence of both indirect and direct linkages. ARACNE evaluates all node triads in a network and removing the edge in the triad with the least evidence of direct control. One restriction of the approach may be the algorithm's removal of all triads in the network, a structure that could be vital in feedback and feed-forward loops (20) (21).

1.3.4 RNA-seq Denoising Methods for Better Gene-Gene Interaction Graph Inference

Chen et al. (90) found that these methods failed to successfully recreate the network for scRNAseq datasets collected through simulations and experiments. Denoising gene-expression patterns may improve the efficacy of these methods. As a consequence, the crucial problem of handling noise and dropout in scRNA-seq profiles is still open. Pre-processing techniques that may reduce noise and sparsity in single cell transcriptomic profiles are necessary for efficient regulation inference.

For scRNA-seq profiles, a few imputation techniques based on cell graphs or employing KNN-based imputation have been presented. MAGIC constructs its affinity matrix in four phases (104). First, a preprocessing step for data such as PCA for scRNA-seq. Then, using an adaptive Gaussian Kernel to convert distances to affinities in such a way that the similarity between two cells decreases exponentially with increasing distance. The probability distribution of switching from one cell to every other cell in the data in a single step is represented by the Markov transition matrix M, which is created by converting the affinity matrix A into the matrix. Finally, exponentiation of M is used to diffuse data, filtering away similarity based on high frequencies, which are often noisy, and increasing similarity based on significant patterns in the data. The imputation stage of MAGIC includes exchanging data across cells in the generated neighbourhoods using matrix multiplication once the affinity matrix has been built $D_{imputed} = M^t * D$.

KNN-Impute (109) is a KNN-based approach that chooses genes with expression patterns like the gene of interest in order to impute missing data. This strategy would identify K more genes whose expression in trials 2-N is most comparable to that of gene A, which has one missing value in experiment 1, and which has a value present in experiment 1 where N is the total number of experiments. The missing value in gene

A is then estimated using a weighted average of the values from the K genes that were closest in experiment 1. Each gene's contribution is weighted in the weighted average based on how closely its expression resembles that of gene A.

Based on the finding that across protocols, the technical noise shown by UMI-filtered scRNA-Seq data closely resembles Poisson statistics, KNN-Smooth was developed (110). Based on partly smoothed and variance-stabilized expression profiles, each cell's closest neighbours are first identified step-by-step, and their transcript counts are then aggregated.

Since our method is also based on KNN based approach, we have performed benchmarking of our approach with other imputation methods for single-cell RNA-seq profiles.

1.4 Graph Based Embedding and Multimodal Integrative Approach for Epigenomic Profiles

High-throughput experimental methods have made it possible to identify and measure biological constituents on a revolutionary scale, from genes and proteins to cells and tissues. Conjointly, these technologies offer a list of components for biological systems such as biochemical pathways, large-scale interaction networks, etc. With the help of network formulations, high-throughput data may be contextualised in a way that makes it easier to comprehend and identify patterns and trends that are important. Networks explain the interactions among several components in a system as determined by scientifically defined or statistically inferred associations, as opposed to a more reductionist approach, which focuses primarily on the function of individual molecular components. As a result, there are many ways to use network information to solve medical problems. For instance, using a priori specified networks, one may include and assess high-dimensional -omics data such as genomics, transcriptomics, proteomics, and metabolomics. Studies are now investigating how to successfully integrate several measurement categories, hence extending the breadth, depth, and accuracy of understanding for biological systems. This is due to the accessibility of varied -omics data and related network analysis techniques. By using network formalisms, such integra-

tive -omics techniques seek to bridge across various data kinds while promoting the development of multiscale networks.

Understanding the genetic determinants of the human body is greatly facilitated by identifying the molecular characteristics that distinguish the types and functionalities of each particular cells. The traditional method of classifying cells is qualitative characterisation, which includes physical appearance, the presence or absence of a few surface proteins, and general cellular activity. Transcriptomic and epigenomic characteristics of cells must be taken into account for a more complete description of cell identity. Numerous high-throughput single-cell sequencing methods were developed recently that analyse the chromatin accessibility, DNA methylation, and gene expression of various individual cells. Together, these data types provide researchers the ability to reevaluate the traditional categorizations of cell types and states in a quantitative, systematic, and objective manner. Such a quantitative definition of cell identity has the potential to fundamentally alter our comprehension of cell biology in a variety of situations, including developmental biology and neurology (22). By providing a reference map of the molecular states of healthy cells, it will be possible to investigate the root causes of cellular irregularity and maybe even design new, specifically targeted therapies. This objective requires an analytical approach that can include multiple single-cell datasets. The "essential" parts of a cell's identity must be identified by such an integrated approach, as opposed to the "dispensable" characteristics that fluctuate depending on the biological context, modality, procedure, or timeframe (39).

1.4.1 Challenges with Single-cell Epigenome Profiles

At the single-cell level, single-cell ATAC-seq (scATACseq) shows changes in chromatin accessibility and may be used to identify the factors controlling cell-to-cell heterogeneity (155) (25). Each diploid genome single-cell open chromatin site, however, contains merely one or two possibilities that can be acquired in scATACseq experiment. Typically, just a few thousand different reads are acquired for each cell, resulting in a large number of authentic open chromatin regions that are devoid of peaks i.e., data signal sequencing. The "missingness" curse and excessive dimensionality are therefore both present in the analysis of scATACseq data.

Single-cell open-chromatin profiles provide specific challenges as compared to the single-cell transcriptome. Currently, single-cell epigenome profiling primarily aims to capture open chromatin regions using DNase I hypersensitive site sequencing (DNase-seq), Micrococcal Nuclease digestion with deep sequencing MNase-seq, or Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) (158) (159) (154). When contrasted to a comparable matrix for a single-cell expression dataset, the read-count matrices constructed using single-cell open-chromatin contain more genomic loci (peaks) as features. The read-count matrices of single-cell open chromatin profiles generated by several research teams often include genomic locations (peaks) that are distinct from one another. Therefore, single-cell open-chromatin profiles cannot be directly analysed using the current algorithms and search techniques established for single-cell expression profiles.

1.4.2 Graph Methods for Integrative Analysis of Multimodal Omics Data

Multimodal technologies that concurrently profile several kinds of molecules inside a single-cell may be used to directly produce multimodal single-cell data, or computational approaches can be used to combine signals from many assays into a single-cell multi-omics dataset. One can find a thorough analysis of approaches for the continuous assessment of single-cell multi-omics in Nathan et. al. (26), Bock et. al. (27), Stuart et. al. (150), Leonavicius et. al. (29) and Zhu et.al. (30).

The fact that all measurements come from the same cells aids in the combination of various modalities obtained through one experiment. The study is limited to using techniques to connect complementary information from several sources, eliminating the requirement to resolve cell identification across modalities. Before testing for association, the correlation of several modalities sometimes begins by summarising the signals of one modality in reference to the genetic entities examined by the other. This strategy is well shown by two studies that incorporated data from the concurrent DNA methylation single cell profiling and gene expression (31) (32). First, the mean binarized methylation rates of genomic regions comprising promoters, gene bodies and enhancers have been assigned to the nearest gene based on 10 kb upstream and downstream windows of the start/end sites of gene transcription. Then, weighted Pearson correlation coefficients

and multi-omics factor analysis (MOFA) (31) (32), an adaptation of group factor analysis to multimodal -omics data, have been used to identify coordinated changes and heterogeneous interactions between methylation states and transcriptional profiles in individual cells. Recent advancements in the MOFA+ approach have increased the capability of MOFA to support large datasets and several batches across modalities (34).

On the other hand, the integration of several data types that are not measured on the same cell needs agglomeration of signals obtained through various protocols and alignment of datasets obtained through various cells. The approach is predicated on the idea that cells belonging to the same type or state have a set of associated characteristics. Such characteristics direct profiles matching across datasets as well as the combining of information from other modalities. Both techniques that precisely orient numerous datasets on the basis of exclusive single-cell dataset and models developed on bulk data may be used to generate signal pairing and dataset integration.

Following the conversion of scATACseq accessibility peaks into gene accessibility based acitivity scores, Conos (152) and Seurat(v3) (150) are specifically being used in the fusion of single cell transcriptomic (scRNAseq) profiles and open chromatin (scATACseq) accessibility profiles. This may be achieved either by adding the signals from the distal and proximal cis-regulatory regions inferred on the basis of scores for co-accessibility, as determined by Cicero (195), or by intersecting the signals from the gene body and an upstream area of a certain length (such as 2 kb). After annotating the methylated areas in accordance with the closest gene, LIGER (39) been applied to combine DNA methylation and gene expression datasets. Clonealign (40), SOMatic (41), and MATCHER (42) have all implemented methods that were especially created in order to incorporate multimodal single cell data.

CONOS (Clustering on network of samples): In Conos (152), a weighted graph representation is created by integrating several datasets, and common populations are found using community detection techniques. The cells of all datasets are used as nodes in the network, which are linked by both intra- and inter-dataset edges. A neighbour mapping approach is used to determine the weights of edges in between datasets in rotation space such as principal components in common space used in MNN (Mutual nearest neighbors). The edges within datasets are scaled down in order to minimize their impact on the merged graph and are weighted based on distances estimated

top principal components within each batch. The graph is then aggregated using either Louvain, Walktrap, or Leiden community discovery methods in order to identify joint clusters connecting cell populations across datasets.

LIGER (Linked inference of genomic experimental relationships): The ideas of the technique based on graph and anchor are used by LIGER (39). To create a latent space where every sample/cell is characterised by various elements unique for each data profiles along with elements common throughout data generated through various protocols, integrative non-negative matrix factorization is initially used in this case. Then, in the factor space, a shared factor neighbourhood network is built, connecting samples/cells along with similar factor loading throughout batches, and Louvain community identification is used to find common clusters across datasets. By matching their quantiles across datasets, the final data alignment created by the joint clusters' factor loadings is adjusted.

Seurat: A unified approach pertaining to transfer learning technique utilizing reference assembly for proteomic, epigenomic, transcriptomic, and spatial single cell profiles is presented by Seurat (150). Even when there are significant technological and/or biological differences across datasets, they may be turned into a common space by identifying pairwise cells comparable between individual cells across datasets, often known as "anchors". As a result, synchronous atlases may be developed at the tissue/organism scale and discrete/continuous data from a reference dataset successfully transferred onto a query dataset.

Integrative approaches' findings for co-embedding of single-cell open-chromatin profiles were not adequate, however it could be because the query set were homogenous and smaller. It could be because integrative approaches like Seurat and LIGER were designed to group homogeneous single-cell epigenome cells query with other dissimilar cells incorrectly since they utilise variation in single-cell data to discover anchor. A vast number of published methods that use canonical correlation (CCA) and principal component analysis (PCA) concentrate on visualising and analysing single-cell epigenome cells amongst a set, which might result in the loss of rare cells which could be only a few cells in the complete data.

1.4.3 Embedding Methods Utlizing Graph and other Techniques for Single-cell Epigenome Profiles

To get an understanding of distinct/mixed cell types or states, it becomes often necessary to analyze and integrate cellular profiles from several sources. There are a few methods which have introduced methods to handle such challenges.

SCALE: In SCALE (Single-Cell ATAC-seq analysis via Latent feature Extraction) (190), the Variational autoencoders framework and GMM (Gaussian Mixture Model) are combined where GMM is probabilistic approach for estimating observed values using a mix of Gaussian distributions. On several distinct datasets produced on various platforms with various methods, and of varying overall data quality, the authors confirm the efficacy of SCALE to extract low dimensional representation that represents distribution of queried scATACseq profiles. Then, SCALE denoised and imputed missing values in the scATACseq profiles and utilized low dimensional features to cluster mixtures of cells into well defined subgroups.

EpiScanpy : The batch corrected k-nearest neighbours (BBKNN) (44) technique is used in EpiScanpy (149) to combine scATACseq datasets generated by various labs and by means of various experimental procedures. BBKNN (batch balanced k-nearest neighbours) provides a hybrid graph consisting of inter,intra-dataset edges which are calculated separately on every batch using K-nearest neighbors and edge weights are given in accordance with the Uniform Manifold Approximation and Projection method.

snapATAC : Harmony (45) is used by snapATAC (46) to incorporate single-cell epigenome profiles from various procedures and labs. Harmony initially clusters cells into many datasets using a PCA embedding. To take into account smoother transition in cell states, it uses soft clustering, which involves putting cells into perhaps many groupings. Instead of establishing separate cell-types, these clusters work as stand-in variables. To prefer clusters having representation across various datasets, it creates a unique soft k-means clustering variation. An information theoretic measure penalises clusters that include substantially unbalanced proprotion of cells in a small batch of datasets. Harmony enables the user to impose a variety of penalties to account for various biological or technical aspects, such as various batches and technological platforms. While preventing local minima that may arise from rapidly maximising representation

over several datasets, the uncertainty represented in the fuzzy clustering retains distinct and continuous topologies and conserve variability. Every data-set has a cluster-specific centroid after clustering, which is employed to calculate the cluster-specific linear correction component. Under ideal circumstances, cell types and cell states are represented by surrogate variables, which are determined by cluster membership. As a result, Harmony's cluster-specific adjustment variables calculated are equivalent with the correction factors unique to each cell type and cell state. Harmony acquires a straightforward linear adjustment function which are responsive to inherent cellular phenotypes in this manner. Finally, a mean of these terms weighted by clusters is allocated to each cell, which is then adjusted by its unique linear factor. As a consequence, susceptible to fuzzy clustering distribution, each cell may have a different correction factor. These four processes are repeated until convergence by harmony. At convergence, further iterations would calculate the same linear correction factors and assign cells to the same clusters.

scFind : scfind is a technique for searching huge databases pertaining single-cell profiles to find cell specific clusters that match specific criteria. Single-cell profiles may be searched using scfind (145) to find sets of housekeeping and cell-type-specific genes. By merging single-cell transcriptome profiles with single-cell ATAC-seq profiles, scfind may be utilised for multi-omics analysis.

Some other RNA-seq embedding methods such as, one where cells are mutually adjacent to one another across batches are known as mutual nearest neighbour (MNN) pairings. Seurat v3 (150), BEER (48), Scanorama (187) and SMNN (47) all make use of the MNN idea.. Other methods can be MINT (188), SCVI (189). The near to twofold character and increasing sparsity in the epigenome profiles, however, may preclude direct applicability of such transcriptomic profile analytical approaches for scATACseq profiles.

Despite the fact that such integration may be highly successful, it is dependent on the data's resolution and the capacity to establish a meaningful connection between gene expression and other modalities. When there is no anticipated biological similarity across datasets, LIGER shouldn't be employed. Additionally, LIGER depends on matching traits that are shared by all datasets being studied. Therefore, it cannot be directly used to combine single-cell metabolomic, morphological, or chromatin con-

formation data with single-cell gene expression data since these characteristics are not clearly related to the expression of specific genes.

Hence, there is a lack of a reliable mapping method that can result in the development of search engines contemplated to accurately find equivalent for scATACseq query profile in large set of numerous single-cell epigenome and transcriptomic profiles regardless of batch effect, despite the availability of large single-cell epigenome atlases (148) (149) (150), as well as approaches for latent space visualisation of these cells.

Therefore, a search engine technique that can handle each individual cell independently of the others would be very helpful in conserving the data of rare and peculiar cells in an experiment and allowing the use of datasets from other studies.

1.5 Protein Biophysical Property Modeling

Amongst the most extensively researched subjects in the field of computational biology is proteins (49). The primary function drivers in a cell are proteins. Proteins have a role in every significant internal process of an organism, from metabolism to defense against invading invaders. Hemoglobin is crucial for delivering oxygen to all regions of the body, and enzymes, which breakdown food, are specialised proteins that carry out assigned functions. A human cell is thought to have between one and three modeling proteins, which demonstrate the abundance of proteins. Thus, knowing proteins is essential to comprehending living things.

Macromolecules known as proteins are composed of lengthy sequences of simple building units. Peptide bonds hold these building pieces, also known as amino acids, together. The amino acids, thus, are often denoted as "residues" in a protein. A plethora of 20 conventional amino acids, and varied arrangements of these 20 amino acids result in various proteins. These amino acids are often represented by one or three English letter codes. Table 1.1 provides a list of these 20 amino acids along with their codes.

The fundamental structure of a protein is pretty routinely represented in a network since it corresponds to a route P_n. However, 3-D structure obtained for proteins is everything along with being monotonous. Protein's linear backbone is folded into an exact three-dimensional shape, with certain protein sections forming helices, sheets, or

Amino Acid	Alanine	Cysteine	Aspartic acid	Glutamic acid	Phenylalanine	Glycine	Histidine	Isoleucine	Lysine	Leucine
3-Letter Code	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu
1-Letter Code	A	C	D	E	F	G	H	I	K	L
Amino Acid	Proline	Asparagine	Methionine	Threonine	Glutamine	Arginine	Serine	Tryptophan	Tyrosine	Valine
3-Letter Code	Pro	Asn	Met	Thr	Gln	Arg	Ser	Trp	Tyr	Val
1-Letter Code	P	N	M	T	Q	R	S	W	Y	V

Table 1.1: Amino acids and their 1,3-letter codes

loops.

Experimental methods for learning about a protein's three-dimensional structure include X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. Protein Databank (PDB), where each protein has a specific PDB identification number (50), receives this information and stores it. One may view this database for free at `http://www.pdb.org`.

1.5.1 Protein Graph Model

A protein's equivalent graph model is known as a protein contact network (PCN). Residues, which are nodes in PCNs, are often connected using various methods. Typically, PCNs are undirected graphs without any self-loops. In the following subsections, the main PCN models that were developed in general are discussed. A PCN model typically starts with a protein's contact map. A contact map is a residue-residue interaction 2-D array of a protein in which a coloured patch at the appropriate cell in the matrix highlights a pair of residues that are in more proximity to one another.

Residue Interaction Network

A simplified protein network model is the residue interaction graph (RIG) model. In a RIG model specific to protein, every vertex constitutes residue, if two residues are adjacent to one another, an edge connects them. The affinity is calculated using their native state structures. Inter-residue distances are computed using per atom's three-dimensional coordinates within a protein (retrieved from Protein Data Bank). To maintain uniformity, the alpha carbon atom is taken into account as the residue's center. If two vertices' proximity, v_1 and v_2 is lower than or the same as a predetermined threshold r_c, then the two vertices are linked, with their associated alpha carbon coordinates

being (x_1, y_1, z_1) and (x_2, y_2, z_2). Overall, we have

$$w_{i,j} = \begin{cases} 1, & \text{if } \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2} \\ 0, & \text{otherwise} \end{cases} \tag{1.25}$$

Here, $w_{i,j}$ is the weighted edge connecting vertices v_i, v_j.

The discussed model produces a straightforward undirected graph. The threshold value r_c is deliberately selected in order to take into account the necessary intensity of attraction that does in fact maintain the stability of the protein structure. The standard threshold, $r_c = 8$Å, is appropriate and relevant as shown by prior publications (210).

RIG with distance information : The RIG model is the best model for mathematically representing a protein. However, it omits the vital detail of the actual separations between residues. A slight modification in the RIG model is also used by using calculated distances as edge weights wherever a RIG edge exists. As a result, we have

$$w_{i,j} = \begin{cases} d_{i,j}, & \text{if } d_{i,j} \leq r_c \\ 0, & \text{otherwise} \end{cases} \tag{1.26}$$

Here $d_{i,j} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$. This produces a weighted, undirected network as opposed to the straightforward RIG model.

Long Range Interactions Graph

The Long-Range Interactions Network is an additional protein graph model (LIN). An associated RIG model's LIN model is a subgraph of that model. Only long-distance interactions—those from the RIG model where the connecting vertices are more than a certain threshold of residues away on the basic structure—are taken into account by LIN. Naturally, the backbone interactions are chosen to maintain network connectivity. In terms of math, we have

$$w_{i,j}^{LIN} = \begin{cases} W_{i,j}, & \text{if } |i-j| = 1 or |i-j| \geq l_c \\ 0, & \text{otherwise} \end{cases} \tag{1.27}$$

where the LIN threshold in terms of residues is represented by l_c. It's important to note that the LIN threshold itself is not a set number; instead, research groups choose somewhere from 6 and 14 amino acids.

Weighted Residue Interaction Graph

There is a lot of research on the RIG and LIN models. However, there are several shortcomings. While the RIG model captures all of the necessary information in terms of the edges, it misses interactions that are far away, which are what truly hold the folded structure in place. It is captured by the LIN model. However, short-range interactions are left out.

A novel protein interaction network model is suggested to address this problem. The shortcomings of the two types mentioned above are handled in this one. It considers all of the RIG model's edges and weights them such that edges that are further apart have larger weighted edges. These weights comprise inversely correlated with spacing in the middle of residues accompanying protein's foundation. Long-range interactions are thus given greater weight. The Weighted RIG (RIGW) model is the name given to this concept. The edge weights are determined as follows:

$$w_{i,j}^{RIG_W} = \begin{cases} |j-i|, & \text{if } d_{i,j} \leq r_c \\ 0, & \text{otherwise} \end{cases} \tag{1.28}$$

1.5.2 Existing Graph-based Methods for Modelling Biophysical Properties of Proteins

A number of efforts contrived to represent protein molecules in form of a residue interaction network of residues related to one another based on their proximity in the protein's 3D structure (210) (52) (53). Protein architectures have been utilised to predict biophysical features by making use of network attributes from graph theory-based metrics of vertices i.e., residues, including clustering coefficient, centrality and betweenness (52). Community network analysis (CNA) is another example, it serves as examination of kinetics of enzymes and complexes of protein/DNA/RNA in order to comprehend their allosteric processes (54) (55). Similar to this, network features based

only on graph-theoretic method have also been used to simulate the protein folding rate. Although such methods emphasise the significance of residue network features, disregarding the biophysical characteristics of amino acids may result in the under utilization of the residues' prior knowledge. As a result, the challenge of how to effectively combine various signals owing the amino acid characteristics and three-dimensional vicinity neighbourhood statistics emerges.

One convenient method to transform a graph which is in the form that resembles a multi-channel image that can be handled by a conventional two dimensional CNN is proposed by Graph-CNN (212). Three phases may be used to summarise the procedure: (1) network node embedding. (2) embedding space compression. (3) repeatedly extracting two-dimensional slices from the compression region and calculating two dimensional histogram for every slice. Stack of network's 2D histograms provides the final "image" representation where each histogram comprises a channel. It should be noted that the final representation of a graph's dimensions is self-contained by graph's innumerable nodes or edges. Images of the same size is used to illustrate both large and smaller graphs.

The following are the ways in which technique overcomes graph kernel limitations are: 1. At the instance level, time complexity is delivered as constant, whereas at the dataset level, time complexity is linear which is achieved by transforming all networks in a certain data-set to depictions of similar dimensions along with utilising traditional two dimensional CNN model to process such graph depictions. Additionally, best node embeddings for particular network may be produced in linear time relative to the number of nodes in network, such as using node2vec. 2. 2D CNN classifier allows for the direct learning of features from the raw data during training, resulting in the highest possible classification accuracy. 3. This strategy makes use of cutting-edge graph node embedding methods that apprehend global as well as local characteristics of networks. It also does away with the need for customized features.

To describe the biophysical characteristics of proteins, graph signal processing (GSP) (231) combines residue attributes and residue network topology. Notably, it demonstrate how the physicochemical characteristics of a protein residue and its structural information affect the folding rate of protein folding. The method also demonstrates the use of regression algorithms to predict three distinct protein characteristics by com-

bining various amino acid residue features with structural data.

Prior research has shown a relationship between the network characteristics of different protein interaction networks and the rate of protein folding (210). These assertions were verified using GSP, which was also utilised to determine if signals on residue, as seen within the frequency spectrum, include explanatory segments that can be attributed to protein's folding rate. A total of fifty two single domain two-state folding proteins were examined to further verify this theory (229). Non-negative signal strength obtained from lower eigenvalues was seen as useful since lower frequencies were thought to be explanatory and high frequencies to be noise (175). This signal's instructive portion is known as the "Low Frequency Component" (LFC)

$$LFC = \sum_{\hat{\lambda_i}=0}^{\hat{\lambda_i} \leq \hat{\lambda_{cutoff}}} |\hat{f}_i| \tag{1.29}$$

In order to find the cutoff frequency that optimises the correlation between LFC and $ln(k_f)$, the cutoff frequency $\hat{\lambda_{cutoff}}$ was altered from 0.01 to $\hat{\lambda_{max}}$ in increments of 0.01.

Deep learning based CNN (Convolutional Neural Network) has attained great enhancement in the area of protein function modelling, still the method requires large pool of data points which can be used to train a CNN architecture. In reality, obtaining a large enough sample size for training is too challenging, and given the limitations of a limited dataset, overfitting is most likely to occur. However, we may utilise the Fourier transform to determine the signal amplitudes needed to duplicate any signal (231). The Fourier transform does have the fundamental restriction that all signal properties must be considered globally. The wavelet transform, on the other hand, depicts the signal in both the time and frequency domain, enabling effective access to the signal's localised information (231).

1.5.3 Effect of Biophysical Properties on Mutation Sites

It has been hypothesised that mutations on conserved residues have detrimental consequences, which may help to elucidate the preservational phenomena which arise underneath strong evolutionary coercion. In fact, the concept has received substantial

prior scientific support (245) (60). Functional residues, however, are not necessarily preserved; frequently, the residues change pertaining a number of grounds, including changing functional precision (61) (62) (63); preserving structural stability (64); and causing allosteric signals such as conformational epitasis (65). As a result, examining variable placements might provide an additional chance to discover functional sites that conservational data has missed.

For the purpose of protein engineering and design, as well as to comprehend protein evolution and genetic diversity in humans, it is crucial to be able to accurately assess and anticipate the impact of amino acid changes in proteins. Natural proteins' sequence analysis and laboratory tests are the two main sources of information on the impact of certain mutations. Phylogenetic studies provide insights on the divergence of protein sequences and the factors that control the eradication of mutations (66) (67) (68) (69). Presuming the mutations' fitness impacts correspond along with the presence in orthologous sequences, protein phylogenies further enable in forecasting if a particular mutation may be neutral or detrimental (70) (71) (72). According to (73) and (74), the algorithms are rather good at predicting mutations that cause diseases. Though several nonsynonymous Single-Nucleotide Polymorphisms (SNPs) anticipated to possess detrimental effect are not distinctly linked to a disease profiles (75), due to their rarity (76) or because a detrimental effect just on one gene frequently has none phenotypic consequences at organism level (77). Actually, there isn't always a correlation between the impact of mutations on isolated proteins and organisms. Additionally, predictors may struggle to attribute harmful consequences to mutations in highly conserved regions which, when altered empirically, seem they have no impact (78). Therefore, comprehensive databases that detail the impacts of every mutation in a gene or protein, irrespective of organismal effects, would significantly enhance prediction (79). To better understand protein evolution, systematic experimental mappings of the effects of mutations within a single gene or protein are therefore crucial. They are also a valuable resource for enhancing predictions (80) (81); and optimising protein design algorithms (82).

Li et al. (213) state that as disease aberrations often occur at residues with high centrality or degree, nearby residues of a mutation site may aid in determining if a mutation is connected to a certain disease. Using network parameters, functional modules in proteins have also been researched. In order to identify functional residues and func-

tional module clusters in rhodopsin that encode typical coevolutionary information in the amino acid network, Park & Kim (214) performed structure-based correlation mutation analysis. A mechanistic understanding of the detrimental effects of amino acid mutations may be gained by connecting their property to the behavior of proteins.

1.6 Objectives and Rationale of thesis

This thesis work addresses the following three main challenges that exist in the field of omics. We also demonstrate how our developed methods can be utilised to understand disease profiles in transcriptomics, epigenomics, and proteomics.

1.6.1 To Denoise Large Read Count Matrices of Single-cell Expression Profiles with Graph Signal Processing for Better Network Inference

In order to improve gene-network inference, a technique is devised to address gene-expression profiles associated noise. Our approach is based on filtering gene expression using graph-wavelets. Our method aims to enhance the performance of current network inference techniques rather than conflict or contend with them. In order to compare different network inference technique output, we did so after processing (Filtered) as well as without processing (Raw) data through graph-wavelet. Several bulk patient samples and single-cell expression samples were used to assess our method. We also looked at how the estimation of the gene network's graph-theoretic characteristics is affected by our denoising technique. It also suggests that inferred interactions are becoming more similar to the real gene-gene interactions when there is an enhancement in intersection between predicted gene-gene networks obtained from two expression profiles for a specific cell-type. Thus, by employing our proposed denoising method prior to measuring the contrast in predicted gene-gene networks owing to ageing or extrinsic stimuli, we may be able to detect real fluctuations in regulatory arrangements. As a result of our tool's ability to filter the single-cell transcriptomic data of pancreatic cells, we could be able to compare the gene-gene networks identified for elderly and juvenile cells. It may be possible to learn more about the susceptibility for disease in the elderly genera-

tion by examining alterations specific to cell-types in gene regulatory networks brought on by aging. Therefore, utilizing scRNA-seq data of juvenile and elderly mouse cells from lung (97), we predicted gene-gene networks for various cell-types. Our strategy of denoising single cell expression patterns increased the consistency of the predicted gene networks for numerous cell-types in juveline and elderly lungs from mice. Another important topic we raised was how the gene regulatory networks of young and aged lung cells vary. We also contrasted the differential expression in lungs infected with COVID-19 with pattern of differences in effect of genes brought on by aging.

1.6.2 To Use Graph-based Integration Method for Embedding Large Single-cell Epigenome Profiles with Different Batch Effects

Our strategy uses cutting-edge computational techniques to compare the single-cell scATACseq profiles that are queried with immense collection of reference single-cell open-chromatin and single-cell expression datasets. With scEpiSearch, one may handle non-identical peak lists from single-cell ATACseq profiles provided by various research teams and figure out the statistical significance of the query's counterpart with single-cell RNAseq and ATACseq profiles. Instead of relying on gene activity as a representative for cell type specificity, scEpiSearch reduces noise and bias among reference profiles by using a gene-enrichment score. Additionally, scEpiSearch overcomes the issue of batch effect, peak-list and species independent query joint-embedding from single-cell ATACseq cells utilising reference cell atlases.

Recently, a number of groups began analysing frozen nuclei obtained from tumour samples using scATACseq. It is well known that tumour cells are heterogeneous and that they often exhibit unidentified intermediate cellular states. In cancer cells, reprogramming and dedifferentiation are often linked to treatment resistance and unanticipated lineage shift. For a clearer apprehension of tumour pathophysiology, it is crucial to contrast single-cell ATAC-seq samples of cancer cells to the large set of existing cells in order to determine their ancestry and potency. Because of this, we initially assessed scEpiSearch's effectiveness in determining lineage using scATACseq samples of K562 cell line and HL60 cell lines as a proof of concept. In order to get a deeper knowledge of regulatory behaviors via their epigenomes, we also utilised scEpiSearch on the epigenomes of embryonic stem cells in this study. We captured heterogeneity, lineage

bias, and stress-response across single-cells.

1.6.3 To Develop an Explainable Predictive Model Using Graph-Wavelet for Modeling Biophysical Properties of Proteins and Measuring Mutational Effects on Diseases

In order to represent the biophysical characteristics of amino acids, we provide a technique based on the graph-wavelet transform of signals of features of amino acids in protein residue networks based on their structures. Additionally, it fared better than approaches based on convolutional neural networks and graph Fourier in predicting the physicochemical characteristics of proteins. Our technique can quantify the influence of each amino acid on the physicochemical characteristics of proteins, even though it cannot anticipate deleterious mutations. Such an assessment of amino acid effects has the ability to elucidate the mechanism behind the detrimental non-synonymous mutation effect. Therefore, for better categorization and deeper comprehension, our technique may highlight patterns of distribution of amino-acid characteristics in the context of a biophysical feature in the structure of the protein.

CHAPTER 2

Denoising Large Read Count Matrices of Single-cell Expression Profiles with Graph Signal Processing for Better Network Inference

2.1 Background

To understand the regulatory mechanisms involved in disease and development, gene regulatory networks are interpreted and used for design and simulation. The network's interdependencies are frequently depicted in form of weighted edges connecting set of nodes, in which the weighted edges potentially depict the regulatory associations between genes. Gene-Gene network can also be employed in the inference of causal models (83), the design and comprehension of perturbation experiments, comparative analysis (84)), and the discovery of new drugs (85). Numerous methods to estimate node interdependencies have been proposed as an outcome of broad applicability of network inference. Majority of methods rely on mutual information, pairwise correlation, or other similarity metrics between values for gene expression that are provided in various conditions or at various times. Furthermore, because of lesser but significant background pattern similarities, resulting edges can be frequently persuaded by indirect dependencies. Even when there is genuine interaction between two nodes, noise, similarity in background patterns, and some other incidental dependencies frequently make it difficult to estimate the effect and strength of the interaction. Therefore, recent methods have begun to infer more confident interactions by using alternative approaches. ARACNE (87), which is a technique that applies statistical cutoff of mutual information or partial correlations by Maetschke et al. (86) could serve as the basis for such an alternative strategy.

Although in a homogeneous population of cells, single-cell RNAseq cells frequently reveal variability within expression scores. This heterogeneity can be used to predict gene regulation networks and uncover dominant cell-type-specific pathways. The best

ways to evaluate gene-gene interaction, however, are still unknown since single-cell RNA-seq profiles are sparse and ambiguous about the distribution of gene expression. As a result, Sknnider et al. (13) recently assessed 17 indicators of association to infer a network based on gene co-expression. They conducted a study and discovered that the two interactions estimation, ϕ (phi) and ρ (rho), performed effectively for estimating gene interactions based on co-expression by utilizing single-cell RNA-seq data. Within a different research (90) the authors independently assessed a few approaches suggested for gene-gene interaction network prediction by utlizing single-cell RNA-seq data, including SCENIC (215), SCODE (216), and PIDC (93). When using single-cell RNAseq cells derived out of simulations/experiments, Chen et al. (90) discovered that these techniques performed poorly in reconstructing the network. Gene-expression profiles can be denoised to increase the performance of such approaches. Therefore, controlling dropout, noise within single-cell RNA-seq profiles may be significantly difficult that still needs to be resolved. There may be biological and technical causes for noise in the single-cell transcriptomic patterns. There could be many possible sources of biological noise in transcriptomic profiles such as some stochastic procedures which could be convoluted in translation or transcription, jagged binding of transcription factors with DNA, expression specific to alleles, Thermal fluctuations etc. Technical noise, however, could result from stochastic detection caused by low RNA levels and amplification bias. The phrase "noise in gene expression" was coined by Raser and O'Shea (95) to describe the degree of heterogeneity in gene expression among cells that should be exactly similar. Raser and O'Shea divided likely causes of disparity in expression of genes into 4 categories:

- An intrinsic stochasticity of biochemical procedure caused by limited count of molecule
- Heterogeneity between cells caused by progression of cell-cycle or some random process like mitochondria partitioning
- Minuscule micro-environmental variations in the tissue
- Genetic mutation

The all-inclusive noise within expression patterns of genes makes it difficult to draw valid conclusions regarding the gene activity regulation in particular cell type. Therefore, a pre-processing technique is needed which is capable of dealing with the sparsity as well as noise in single-cell RNAseq cells to enable accurate regulation prediction.

Using techniques taken from graph theory, the anticipated gene-gene interaction network can be additionally examined such as to deduce important regulatory patterns in each specific cell-type. Finding communities and modules of genes (84), as well as calculating gene relevance in terms of centrality, are typical downstream analytic techniques. Similar to gene-expression profiles, inferred gene networks can be used to compare two groups of cells and identify differences in the regulatory pattern brought on by illness, exposure to the environment, or aging. Recent interest has focused on a comparison of regulatory changes brought on by aging, in particular since older people have a higher chance of metabolic disorders and mortality on the basis of infections. Researchers are concerned about this issue, particularly given the current condition of pandemics caused by new COVID-19 (SARS-COV-2), where elder people have a greater likelihood of mortality. Why do elderly cells from lung hold a greater chance of onset of severe COVID-19 contagion is major question. However, utilising gene-gene interaction network prediction with erratic single-cell RNAseq cells from lung to understand regulatory alterations brought on by ageing is not a simple task. Thus, a method that can suppress batch effect, noise is required for network biology-based examination of single-cell RNAseq dataset of lung aging cells (97).

Overall, this objective showcased the effectiveness of the graph-wavelet-driven (which is a graph based method) gene-expression filtering approach in enhancing gene-network inference. Moreover, by applying this method to examine gene expression in the context of aging and COVID-19 infection, we shed light on the specific gene regulatory networks associated with these conditions and hence, helps decipher omics signature of diseases. This contributes to our understanding of the underlying biological processes and provides valuable insights for further investigation and potential therapeutic interventions.

2.2 Materials and Methods

As defined by various researchers, including Raser and O'Shea (95), we used the phrase noise in gene expression to direct towards computed degree of fluctuation in gene expression across cells that should be identical. As a result, we started by creating a base-graph (networks), which consists of edges connecting cells that are purportedly

identical. With the use of such graph as base and applying graph-wavelet transform, we can measure how much each gene's expression varies between connected samples at various graph spectral resolutions in each sample (cell). To achieve this, we first determined the distances between samples (cells). Principal component analysis or any dimensionality reduction techniques for expression data can be used to obtain a more accurate estimation of the distances between samples (cells). Each cell/sample was treated as vertex of the network. Two vertices were only linked with edges provided one of those nodes would be one from remaining cells K nearest neighbours. On the basis of number of sample/cell in expression profiles, value of K here was chosen, which is in the range of 10 to 50. As a result, we used K-nearest neighbours (KNN) to generate the primary adjacency matrix by making use of the euclidean distance scores among cells in expression matrix. The computed adjacency matrix was used such that we can create base graph. As a result, every node within base-graph represents a different cell, edge weights representing their Euclidean distance from one another.

A collection of edges E representing interconnections among cells (if present) and weighted function that produces positive weighted associations among cells constitute the weighted graph G that was created using a KNN-based adjacency matrix of cells. Such weighted matrix may be alternatively described in form of NXN (where N is the number of samples/cells) weighted adjacency matrix A, here A_{ij} would be zero if cells i and j do not have any edge between them, and $A_{ij} = weight(i, j)$ otherwise.

Total edge weight that are incident on a particular cell in the graph constitutes the cell's degree. Additionally, the degree (diagonal) matrix D of such network would consist degree $d(i)\ if\ i = j$ and degree 0 otherwise. L is the unnormalized graph Laplacian operator, and it can be defined such that it would be equal to $D - A$ for a graph. This is the definition of the normalised version of the graph Laplacian operator.

$$L^{norm} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}AD^{-1/2} \tag{2.1}$$

According to sandryhaila et. al. (131), the two separate laplacian operators generate two distinct sets of eigenvectors. On the other hand, for our approach of graph derivation between cells we have used normalised version of the Laplacian operator. Further applications of graph Laplacian include signals' Graph Fourier transformation logged on every node ((218) (131)).

For fourier transform coefficients filtering, Chebyshev-filter was used for filtering single-cell RNAseq in Fourier domain. We projected the raw graph (where each node refers to a different cell) object (101) using the expression of each gene, using it as a signal. We started with the signal's forward Fourier transform, applied a Chebyshev filter, and took inverse transform of signal to get the denoised profiles. For each gene, the same procedures were repeated. Thus,We ultimately get filtered gene expression as a result which is used to further get insights.

2.2.1 Spectral Graph Wavelet Transform

Approach similar to fourier transform, spectral graph wavelet is implemented which requires the selection of a positive real-valued kernel function that may act in form of bandpass filter. The wavelet operator provided by the graph laplacian's rescaled kernel function ultimately yields graph wavelet coefficients at every scale. On the basis of the spectral representation of network, a self adjoint operator function using continuous functional calculus can be created. Although this may be accomplished by utlizing eigenvalues, eigenvectors of the laplacian L, (218), for the graph with finite dimensional Laplacian. $g = g(L)$ gives the wavelet operator. A signals' wavelet coefficients with $scale = 1$ are provided by $T_g f$. When applied to eigenvectors U_l, this operator acts as

$$T_g U_l = g(\lambda_l) U_l \tag{2.2}$$

As a result, operator T_g modifies each graph Fourier coefficient by operating on every graph signal.

$$\widehat{T_g f(l)} = g(\lambda_l)\hat{f}(l) \tag{2.3}$$

Further, Inverse Fourier transform can be defined

$$(T_g f)(m) = \sum_{l=0}^{N-1} \hat{f}(l) U_l(m) \tag{2.4}$$

Wavelet operator is interpreted as $T_g^s = g(sL)$ for each scale s. When these wavelet operators are applied to δ_n, which yields signal with value 1 on node n else 0 and are localized to yield individual wavelets (218). Consequently, if all scales are taken into account, the inverse Fourier transform is then used to calculate wavelet coefficient using

the concept of convolution which is calculated as

$$\psi_{t,n}(m) = (t_g^s f)(n) = \sum_{l=0}^{N-1} g(s\lambda_l)\hat{f}(l)U_l(n) \tag{2.5}$$

Despite using Fourier domain filtering in this case, we measured the wavelet coefficients of every gene expression signal on various scales. Each scale's wavelet coefficients were filtered using thresholding. On wavelet coefficients, we used both hard and soft thresholding. However, ee used the conventional techniques Bayes Shrink, Sure Shrink to implement soft thresholding.

2.2.2 Selecting Optimal Threshold for Graph-wavelet Coefficients

It has been a focus of intense study to determine the best wavelet coefficient threshold for noise removal from linear signals and images. We investigated information theoretic criteria i.e. minimal description length principle (MDL) and assessed both soft and hard thresholding methods. The user of our application, GWNet, may choose from a variety of threshold-finding alternatives, including visuShrink, sureShrink, and MDL.

As appropriate soft-thresholding with respect to Graph-wavelet coefficient is fundamentally a subject of much investigation and might require additional fine-tuning, hence hard-thresholding was chosen for majority of the data used in this study.

Another option is to employ a hard-threshold criteria on the basis of the best protein-protein interaction and projected gene-network overlap (PPI). We discovered that cutoff determined through MDL criterion as well as the most overlap between the projected network and the ground set interactions/PPI were found to be in between 60–70 percentile when applying it to various data sets. We need a consistent percentile threshold to cutoff graph-wavelet coefficients so that projected networks from different data sources can be compared. In order to analyse various data sets uniformly, We established a 70 percentile baseline cutoff value. As a result, wavelet coefficients with absolute values below the 70th percentile were set to zero in default mode.

2.2.3 Methodologies for Inferring Gene Interaction Network

Any network inferences technique may be fed into the adaptable GWNet tool to make regulatory predictions utilizing graph-theoretic methodology. Thus, gene expression raw counts in form of fragments per kilobase of exon model per million reads mapped (FPKM) for single-cell RNAseq profiles were used. The quantile normalisation and log transformation were employed in order to pre-process single-cell gene expression. To begin with, we simply estimated the measure of gene interdependencies using the Spearman and Pearson correlation. Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) was also utilised for inferring the gene regulatory network. For each gene-pair, ARACNE first calculates the mutual information. Then, all potential gene triplets are taken into account, and indirect interactions were removed applying Data Processing Inequality (DPI) method. The following inequality is supported by DPI provided gene i, gene j exhibit reliance on gene k rather than direct interaction.

$$I(G_i, G_j) \leq min(I(G_i, G_k), I(G_j, G_k)) \tag{2.6}$$

Here, i and j genes' mutual information is represented by the symbol $I(G_i, G_j)$. ARACNE also eliminates interactions with mutual information (MI) that is shared less often than a predetermined eps parameter.

For predicting gene-gene co-expression networks obtained from single-cell RNAseq profiles, Skinnider et al. have shown the precedence of 2 proportionality metrics, $rho(\rho)$, $phi(\phi_s)$ (140). As a result, we also assessed the advantages of gene-expression denoising based on graph-wavelet using the proportionality metrics ρ , ϕ_s.

Proportionality metric ϕ is described below

$$\phi(G_i, G_j) = \frac{var(G_i - G_j)}{var(G_i)} \tag{2.7}$$

Where var() is the variance function and G_i would be array carrying the i genes' expression log scores over numerous samples/cells. The symmetric representation of ϕ is denoted by

$$\phi_s(G_i, G_j) = \frac{var(G_i - G_j)}{var(G_i) + var(G_j)} \tag{2.8}$$

While rho is described below

$$rho(G_i, G_j) = 1 - \frac{var(G_i - G_j)}{var(G_i + G_j)} \tag{2.9}$$

We used the 'propr' package 2.0 in R to estimate the ρ and ϕ metrics of proportionality. (141).

2.2.4 Raw-Filtered Gene Network Comparison

Ground truth was compared with the networks estimated from filtered and raw gene expression profiles. In order to assemble the ground truth for single-cell profiles, thus, Human Integrated Protein-Protein Interaction Reference (HIPPIE) repository (142) was used. However, the known interaction set for the DREAM5 challenge expression set was available at prior. We took into account all potential network edges and ordered them according to their importance as per edge weights. Through comparison to edges in the known set of interactions, we were able to determine the area under Receiver operator curve for filtered as well as raw interactions. Receiver operator is common assessment statistic in the field of machine learning that was modified for use in the DREAM5 evaluation approach. Here, we modified the Receiver operating curve by using a certain number of edges ordered by weights plotted on the X-axis rather than the false-positive rate. For assessment, all plausible edges obtained from gene-gene interactions inferred from filtered as well as raw graphs are collected and sorted according to their weights in the network. By comparing the fold change between the raw and filtered scores, improvement was determined.

2.2.5 Comparison with Other Methods

We contrasted our methodology outcomes obtained by denoising using graph-wavelet with those of additional techniques designed for imputation or noise reduction in scRNA-seq profiles. We also utilised Graph-Fourier based filtering for comparison (101), MAGIC (104), scImpute (105), DCA (106) , SAVER (107), Randomly (108), KNN-impute (109).

MAGIC

High dimensional scRNA-seq data are imputed using Markov Affinity-based Graph Imputation of Cells(MAGIC). MAGIC builds graph to get smoothed characteristics and then restores the data's organisational structure. By detecting the most similar cells, it aggregates highly similar cells to create a graph. Imputation is used to adjust for dropout and noise in this way. By building a weighted affinity matrix, it employs data diffusion. The Markov transition matrix illustrates the probability distribution of further transitions between cells. The URL[1] was utilized to install the method. With the principal component (PC) parameter setting of 20, parameter t setting of 6 for the power of Markov affinity matrix, autotune parameter setting of 10, number of closest neighbours setting of 30, and scaling setting of 99^{th} percentile, we employed MAGIC.

ScImpute

ScImpute begins with dimensionality-reduction (PCA) on the expression matrix before calculating the distance matrix. It employs a first quartile and third quartile for the set of neighbours to eliminate outlier cells. On the remaining cells, spectral clustering is carried out. ScImpute employs a statistical model to determine whether or not a dropout results in a zero value. When dropout occurs, it is assumed that genes exhibit a bimodal pattern of expression, which may be explained by two-component mixture models. Gamma distribution and normal distribution, which account for dropouts and reflect the actual amount of gene expression, are the two components. Two gene sets are created, the first of which includes genes that need imputation and have a certain threshold for gene dropout probability in a cell, and the second of which includes genes that do not have exact gene expression. Genes from the second group are used to gauge how similar cells are to one another. Additionally, using non-negative least squares regression, data from the second gene set's expression in other cells that are comparable is used to impute the first gene set's genes. The URL for scImpute was [2]. We also used the default ScImpute settings, which were : threshold set at 0.5 for dropout probability and 2 for cell subpopulation.

[1] `https://github.com/pkathail/magic`
[2] `https://github.com/Vivianstats/scImpute`

DCA

By taking into consideration the count pattern, sparsity, and overdispersed nature of scRNA-seq data, deep count autoencode (DCA) employs deep autoencoder with loss function of zero-inflated negative binomial (ZINB), eliminating dropouts and noise in read-count. For each gene's input expression data, DCA utilises an autoencoder model to learn the dropout (π), dispersion (θ), and mean (μ) parameters of the ZINB distribution. The size of the input layer and the three output layers, which provide the aforementioned three parameters, have the same number of genes in this autoencoder network. But unlike a conventional autoencoder, DCA contains three output layers, each of which represents the gene-specific parameters and as a result, each gene has a loss function that compares its output to its original input. As a result, the output layer may be thought of as ZINB regression with new representations of cells as predictors. From the following URL: [3], DCA was downloaded. For optimization, we utilised DCA with default settings and RMSProp with 32 batches, 20 epochs, and a learning rate of 0.001.

SAVER

Using regularised regression prediction and the empirical Bayes approach, Single-cell Analysis Via Expression Recovery (SAVER) attempts to recover real gene expression patterns from sparse and noisy scRNA-seq data. A multi-gene prediction model based on adaptive shrinkage is hence called SAVER. SAVER employs the poisson distribution to manage cellular biological variance. In order to take into consideration the prediction uncertainty, it assumes a gamma prior with a set of mean to prediction and parameter dispersion. The parameter dispersion gauges how well the prediction mean can forecast a gene's expression. SAVER recovers gene expressions using posterior mean. The URL for SAVER is [4]. SAVER was also run using default settings, with a maximum of 300 genes allowed for prediction, 50 for lambda value used for cross-validation, and 5 folds in cross-validation.

[3] `https://github.com/theislab/dca`
[4] `https://github.com/mohuangx/SAVER`

Randomly

The technique Randomly use Random matrix theory to denoise single-cell data. Creates a wishart matrix at random and uses singular value decomposition to calculate the whole collection of eigenvalues and eigenvectors for the creation of wishart matrix. Using eigenvalues that are higher than the Tracy-Widom critical threshold, biological signals are extracted. The genes with the highest variation between eigenvector signals and eigenvector noise signals are then chosen. As a result, the method ignores genes with a significant variation over the eigenvector signal. From the github URL : [5]), Randomly was downloaded. Additionally, randomly was used with the default values of 0 for min_tp, $min\_genes\_per\_cell$, and $min\_cells\_per\_gene$.

KNN-Impute

KNN-impute uses the k-Nearest Neighbour method to accomplish the imputation of missing values in the expression matrix. The R package bnstruct contains the KNN-Imputed function knn.impute. We selected 10 as the value of k for KNN-Imputed.

Graph Fourier Transform

Typically, the Fourier transform magnifies a signal into the Fourier basis of filters-invariant signals. A conventional Fourier transform is therefore provided by

$$\hat{f}(\omega) = \int e^{-it\omega} f(t)\, dx \tag{2.10}$$

Complex exponentials are the eigenfunctions of the one-dimensional Laplacian in this case, $e^{(-it\omega)}$. As a result, the Fourier transform or calculating the Fourier coefficient $\hat{f}(\omega)$ is defined as the inner product of the signal $f(t)$ and the Eigen function of the Laplacian. The formula for inverse Fourier transform is

$$\hat{f}(t) = \frac{1}{2\pi} \int \hat{f}(\omega) e^{-it\omega}\, d\omega \tag{2.11}$$

where the signal $f(t)$ is expanded as a weighted sum of the Laplacian eigen func-

[5] `https://github.com/RabadanLab/randomly`

tions. By substituting graph Laplacian for the Laplacian in the previous equation, one may also obtain the graph Fourier transform in a similar manner. The orthonormal eigenvectors of the symmetric Laplacian matrix are designated as U_l for $0...N-1$. Because of their symmetry, eigenvalues may be sorted in ascending order and are likely to be real. Additionally, the minimum eigenvalue for the graph Laplacian is 0, and the eigenvalues are all non-negative. Eigenvalues λ must fulfil $0 = \lambda 0 < \lambda 1 \leq \lambda 2... \leq \lambda(N-1)$ given that graph G is connected. Consequently, Fourier transform would be for every signal graph :

$$\hat{f}(l) = \sum_{n=1}^{N} U_l^*(n) f(n) \tag{2.12}$$

and inverse graph Fourier transform [3] as

$$f(n) = \sum_{l=0}^{N-1} \hat{f}(l) U_l(n) \tag{2.13}$$

Chebyshev-filter was utilised for gene expression profile filtering in the Fourier domain. This strategy does not include this technique. We projected the raw graph object (where each vertex corresponds to a different sample) with the expression of each gene at a time, using it as a signal. We started with the signal's forward Fourier transform, applied a Chebyshev filter in the Fourier domain, and then inverse-transformed the signal to get the filtered expression. For each gene, the same steps were repeated. We would ultimately get filtered gene expression as a result. With beta set to 2, offset set to 0, and order set to 5, the Chebyshev filter was created for filtering in the Fourier domain.

2.2.6 Differential Centrality

Centrality is a key concept in graph theory and network-based research for identifying significant nodes in a graph. It measures how central node is or how significant a node in the network is. The majorly applied centrality measures are closeness, degree, betweenness, PageRank, and eigenvector centrality, which further offer crucial analytical information about the network and the nodes, and may be defined in a variety of ways. To determine differential centrality in ageing data, we have employed two centrality metrics, namely degree and PageRank.

2.2.7 Data Sources

The DREAM5 portal's bulk gene-expression data was downloaded for the initial evaluation of the method ([6]). From the GEO database, the single-cell expression profile of mESC produced using several procedures, (102), was downloaded from (GEO id: GSE75790). From the GEO database, single-cell expression profiles of pancreatic cells from people of various ages were retrieved from (GEO id:GSE81547). The GEO id for the scRNA-seq profile of murine ageing lung from Kimmel et al. (97) is GSE132901. While the GEO id: GSE132901 is available for the ageing lung scRNA-seq data published by Angelids et al.

2.3 Results

Our approach is based on the idea that cells (samples) with similar characteristics will have more similar gene expression profiles. As a result, we start by creating a network where two cells are linked by an edge if one of them is one of the other's top K closest neighbours (KNN). Following the construction of a KNN-based network between cells (samples), we filter the expression of one gene at a time using a graph-wavelet technique (Figure 2.1). We employ a gene's expression as a signal on the nodes of the cell graph. In order to conduct spectral decomposition of the graph signal, we use a graph-wavelet transform. Following the graph-wavelet transformation, the wavelet coefficient threshold is selected using sureShrink and BayesShrink or a preset percentile value defined following extensive testing on several data sets. To rebuild a filtered expression matrix for the gene, we employ the coefficient for inverse graph-wavelet transformation's preserved values. For gene-network inference and other downstream analyses of regulatory differences, filtered gene-expression is employed. We evaluated inter-dependencies across genes for assessment purposes using five distinct co-expression metrics, including Pearson and Spearman correlations, phi and $rhoscores, and ARACNE$.

[6] `http://dreamchallenges.org/project/dream-5-network-inference-challenge/`

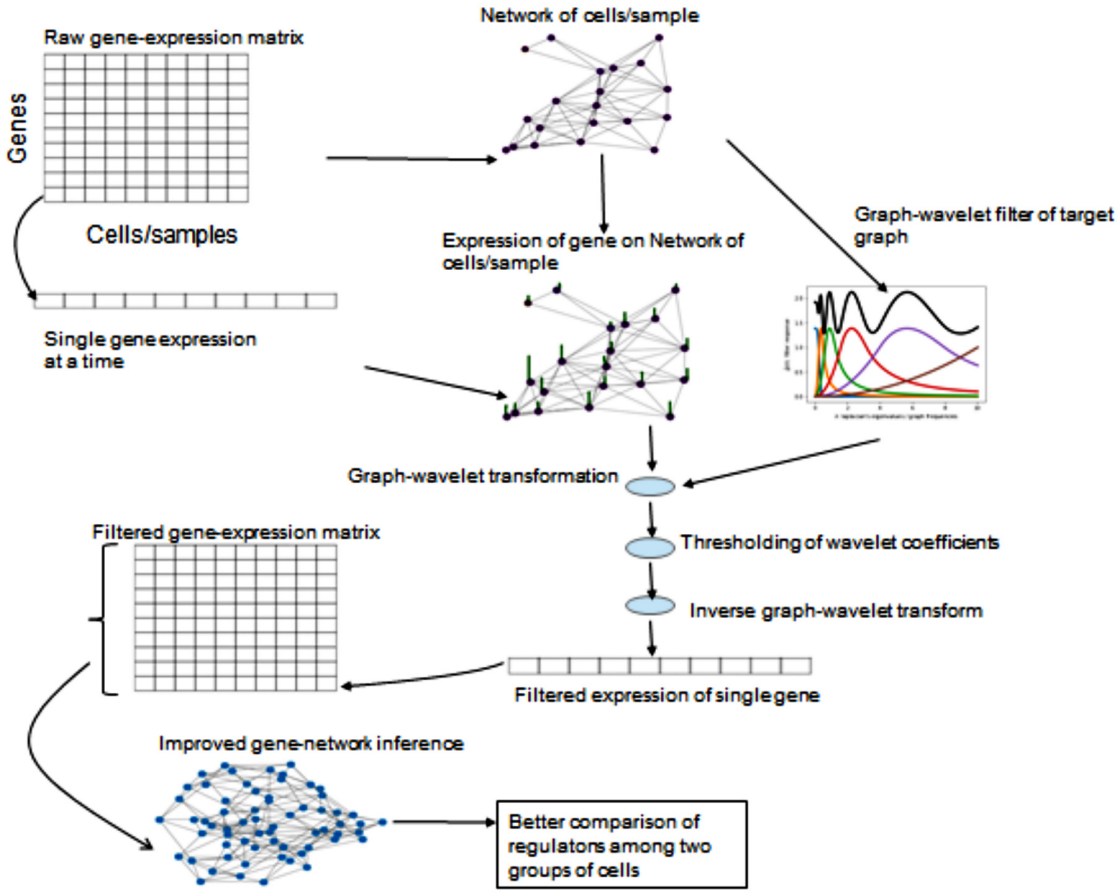

Figure 2.1: The GWNet schematic flow. The first step is creating a KNN-based network connecting samples. For the KNN-based network of samples, a graph wavelet based filter is learnt. Graph-wavelet transform is used to filter the gene-expression of a single gene at a time. For network inference, filtered gene-expression profiles is utilized. Calculations of differential centrality between cell groups are done using the inferred network. (Mishra et. al. (247))

2.3.1 Assessment using DREAM Challenge Bulk Expression Profiles

A bulk sample expression profile may include both biological and technological noise ((95)). We initially assessed the effectiveness of our technique on a bulk expression data-set in order to test the premise that graph-based denoising may enhance gene-network inference. We utilised four data sets that the DREAM5 challenge consortium made accessible (99). Based on the original expression profiles of the single-celled eukaryotes Saccharomyces cerevisiae and S aureus as well as the bacteria Escherichia coli, three data sets were created. While GeneNetWeaver, which uses the chemical

Langevin equation to simulate molecular noise in transcription and translation, was used to replicate the fourth data set using an in silico network (100).

For each of the four data sets, the real positive interactions are also accessible. Utilizing three alternative methods to cutoff the wavelet-coefficients, we contrasted graph-wavelet based denoising with graph-fourier based low pass filtering. With correlation, ARACNE, and rho-based network prediction, we were able to improve the score by 5 to 25 percent compared to the raw data using the DREAM5 criteria (99). In three of the four DREAM5 data sets, the gene-network prediction method based on phi_s showed improvement (Fig. 2.2). Following the graph-wavelet based denoising of the simulated data (in silico) from the DREAM5 consortium, all five network inference algorithms shown improvement (Fig. 2.2). Further, graph-wavelet based filtering performed better than Chebyshev filter-based low pass in Fourier domain filtering. It draws attention to the fact that even large-scale bulk gene-expression sample data may include noise and that denoising it using a graph-wavelet after creating a KNN-based graph among samples has the potential to improve inference of gene networks. It also emphasises another well-known fact in the domain of signal processing—namely, that wavelet-based filtering is more adaptable than low pass filtering.

2.3.2 Gene-networks Inference Improved by Graph-wavelet based Denoising of Single-cell Expression Profiles

In single-cell expression profiles, noise and dropout are more apparent than in bulk data. Technical issues prevent actual expression from being detected, which leads to dropouts. Since low-pass filtering fills in the background signal at missing values and suppresses high-frequency outlier signals, using it after graph Fourier transform seems to be an obvious choice (101). However, a blind smoothing of a signal may not be successful in the absence of knowledge regarding cell-type and cell-states. Therefore, to analyse the gene-expression data set from the scRNA-seq profile, we used graph-wavelet based filtering.

We started by using the Ziegenhain 2017 scRNA-seq data set of mouse embryonic stem cells (mESCs). We utilised gene regulatory interactions documented by another research team (98) in order to assess network inference objectively. Our method of pre-

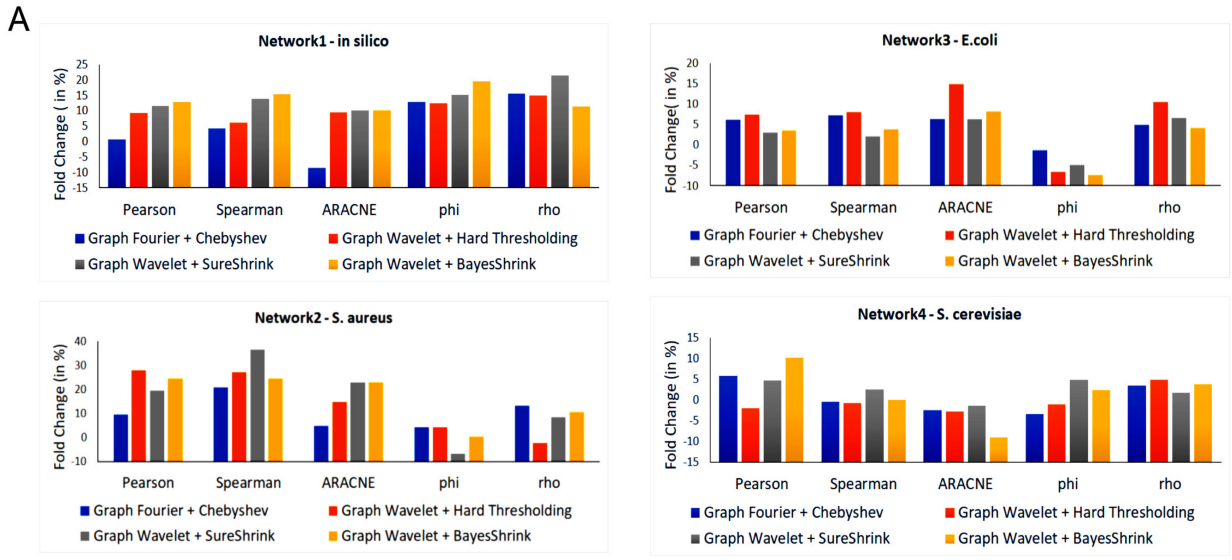

Figure 2.2: Gene-network inference is improved by graph-wavelet based gene-expression denoising. (A) Network inference techniques' performance utilising DREAM5 challenge's bulk gene-expression data sets. Low pass filtering based on the graph Fourier transform was compared to three distinct methods of graph-wavelet coefficient thresholding. The Y-axis displays the fold change in the receiver operating characteristic curve's (ROC) area under the curve (AUC) for the overlap of the anticipated network with the ground truth set of interactions. The default value of 70% percentile was applied for the hard threshold. (Mishra et. al. (247))

processing the mESC scRNA-seq data set using graph-wavelet technology increased the efficiency of gene-network inference techniques by 8–10%. (Fig. 2.3). The gold-set of interactions, however, is often limited and restricts a fair evaluation of performance in gene-network inference. As a result, we also adopted a different strategy to evaluate our methodology. We utilized a measure of network overlap that was derived from two sets of scRNA-seq data that were of the same cell type but had distinct technical biases and batch effects. The inferred networks from both data sets will exhibit substantial overlap if they are more closely related to the real gene-interaction model. We utilized two sets of mESC scRNA-seq data produced using two separate protocols for this purpose (SMARTseq and Drop-seq).

We also employed a few additional imputation and denoising approaches that were reported to filter and estimate the missing expression values in scRNA-seq profiles in order to compare consistency and performance of our method. We tested 7 other algorithms, including Graph-Fourier based filtering (101), MAGIC (104), scImpute (105), DCA (106), SAVER (107), Randomly (108), and KNN-impute (109). The overlap of

the projected network with the known interaction was improved more by graph-wavelet based denoising than by the other 7 methods introduced for imputing and filtering scRNA-seq profiles (Figure 2.3). The other 7 approaches did not significantly enhance the AUC for overlap among the gene-network inferred by two sets of mESC data, similar to the graph-wavelet based denoising (Fig. 2.3). Even though they were denoised individually, graph wavelet-based filtering improved the overlap significantly across networks estimated from various batches of the scRNA-seq profile of the mESC (Fig. 2.3B). Due to graph-wavelet based denoising, the overlap between projected gene networks increased by 80% using ϕ_s based edge weights (Fig. 2.3B). The improvement in network overlap between the two batches suggests that graph-wavelet denoising differs from imputation approaches and has the potential to significantly enhance inferring gene networks from expression data.

2.3.3 Age-based Regulatory Differences Shown by Improved Gene-network Inference from Single-cell Profiles.

Improved overlap between inferred gene networks from two sets of expression data for a particular cell type also suggests that predicted networks after denoising are more similar to true gene-interaction patterns. Thus, by employing our denoising method before measuring the difference in predicted gene-networks owing to ageing or external stimuli, we may be able to detect real changes in the regulatory pattern. We used the scRNA-seq profile of young and old pancreatic cells, which had been filtered by our method (111), to compare the gene networks identified for each group. According to Martin et al., there were three age categories: young adult (21–22 years), aged (38–54 years), and juvenile (1 month–6 years). We independently used graph-wavelet based denoising on three distinct sets of pancreatic cells. In other words, while denoising, we did not mix cells from various age groups. A single-cell profile of pancreatic cells that had been denoised using graph-wavelets performed better in terms of overlap with protein-protein interactions (PPI) (Fig. 2.4A). Although we employed PPI to gauge refinement in gene-network inference, similar to Chen et al. (35), it's possible that not all gene interactions were taken into account. As a result, we also evaluated our technique for scRNA-seq profiles of pancreatic cells using the criterion of an increase in overlap across projected networks for the same cell types. Denoising scRNA-seq profiles

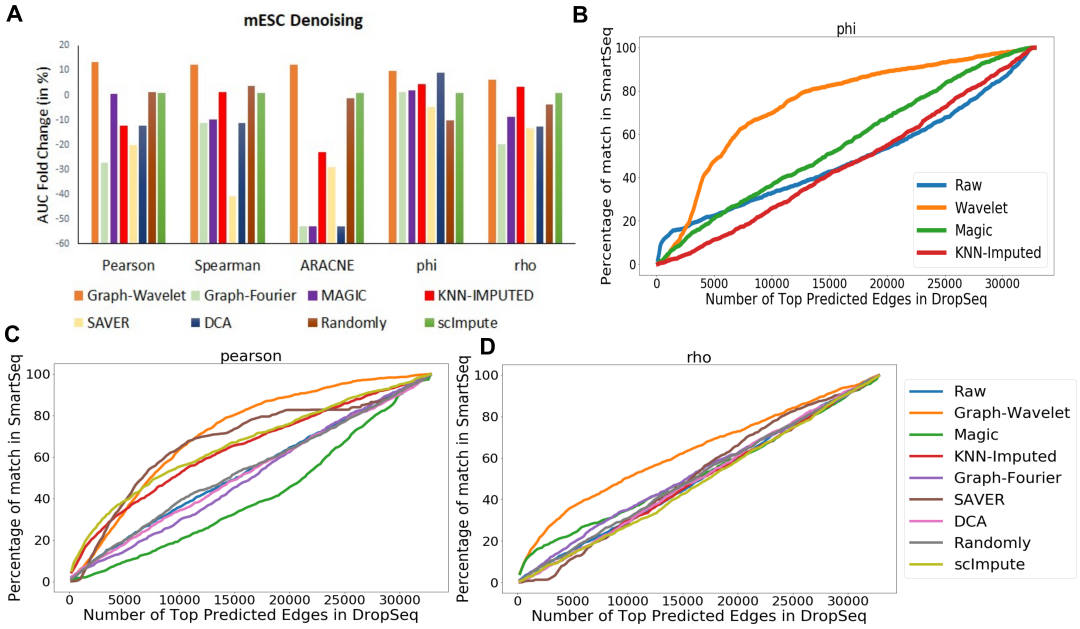

Figure 2.3: (A) A comparison of a few smoothing and imputing techniques with graph-wavelet based denoising. Fold change in the AUC-ROC for the predicted gene network's overlap with a set of known interactions For mouse embryonic stem cells, the gene-networks were predicted after the imputation or filtering of scRNA-seq profiles using various techniques. The interactions in the gold set were taken from (103). (B) A comparison of the consistency of the gene-interaction network prediction using denoising using graph wavelet with other relevant smoothing and imputing techniques. Here, the Phi (ϕ_s) score was used to forecast the gene regulatory network. To test resilience against the batch effect, inferred networks from two scRNA-seq profiles of mESC were compared. (C, D) Reliability in the gene-interaction network prediction for mESC using two batches of scRNA-seq data. Results for Pearson correlation and a co-expression network based on ρ scores are shown below. (Mishra et. al. (247))

significantly enhanced overlap in estimated gene networks between elderly (aged) and young people's pancreatic cells (Fig. 2.4B). In order to put the three age groups on a level basis and measure the variance of expression across cells for each gene, we quantile normalised the original and denoised expression matrix. The median variance of gene expression was larger in elderly and young pancreatic alpha cells than in juvenile cells. The variation level of genes across the three age groups, however, became almost equal and had a comparable median value following graph-wavelet based denoising (Fig. 2.4C). It should be noted that estimating the proportion of variations attributable to transcriptional or technological noise is not simple.

Transcriptional bursting happens when a gene promoter cycles between a "on" and a "off" state for varying amounts of time. Every time the promoter switches to a "on"

state, transcription is produced in bursts. The frequency, duration, and amplitude of bursts affect how much RNA is produced from a particular gene (144). This is a result of a technique called transcriptional bursting. This noise is often described in terms of the degree of gene expression variability seen within a population of cells.

However, single-cell expression patterns of elderly and young people seemed to have less noise after graph-wavelet based denoising. Studying variations in the effect of genes has been done using differential centrality in the co-expression network. However, noise in single-cell expression patterns might result in fictitious centrality variations. As a result, we demonstrated the differential gene expression levels in the network that was inferred from the scRNA-seq profiles of young and old cells. When compared to the denoised version, the networks inferred from unfiltered expression showed a much greater number of non-zero differential degree values (Fig. 2.4D). Thus, denoising seems to lessen centrality differences, which may be caused by noise's randomness. The characteristics of the genes whose variance decreased the greatest as a result of graph-wavelet based denoising were then examined. Surprisingly, we discovered that diabetes mellitus and hyperinsulinism were substantially linked with the top 500 genes with the larger drop in variance caused by denoising in elderly pancreatic beta cells. Conversely, the top 500 genes in young pancreatic beta cells with the largest variance decline showed either no or a negligible connection with diabetes (Fig. 2.4E). Pancreatic alpha cells showed the same pattern. Such a finding suggests that ageing increases the stochasticity of the level of gene expression connected with pancreatic function, and denoising may aid in correctly revealing these genes interdependencies with other genes.

2.3.4 Improved Gene-network Inference Utilized to Analyze Regulatory Variations Between Young and Aged Lung Cells.

It may be possible to learn more about the susceptibility for illness in the older population by examining cell-type-specific alterations in regulatory networks brought on by ageing. Therefore, using scRNA-seq profiles of young and elderly mouse lung cells reported by Kimmel et al. (97), we inferred gene networks for various cell types. Numerous cell types, including bronchial epithelial and alveolar epithelial cells, fibroblast, alveolar macrophages, endothelial, and other immune cells, make up the lower lung

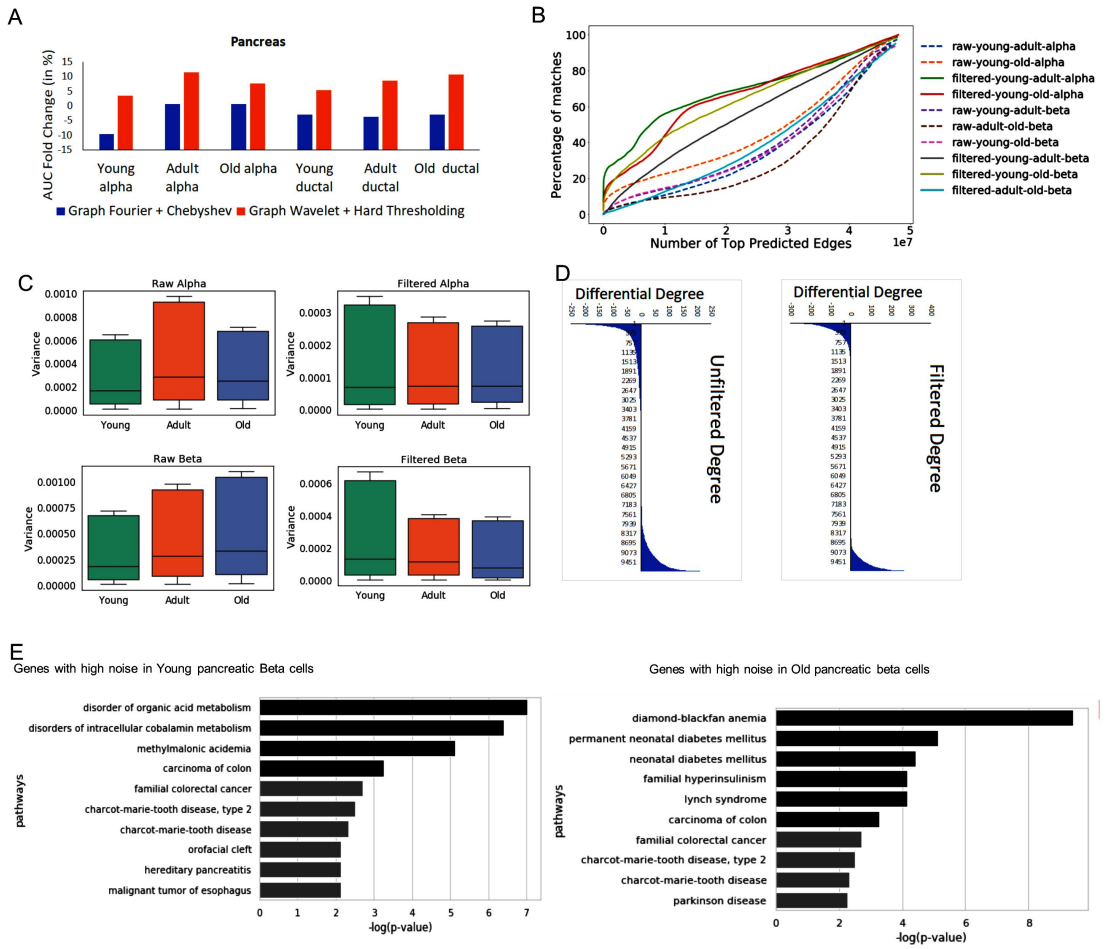

Figure 2.4: Performance and noise analysis for pancreatic single-cell RNA-seq profile.(A) Performance as measured by the estimated network's overlap with the data set on protein-protein interactions. (B) A consistency assessment of the anticipated network. It's crucial to minimise noise-related discrepancies when comparing two networks. The results for the correlation-based co-expression network are shown below. (C) This figure illustrates the differences in gene expression across single cells both before and after denoising (filtering). Gene variation in a cell type was determined independently for each of the three phases of aging (young, adult and old). Compared to young alpha and beta cells, the variance (estimated noise) is larger in older alpha and beta cells. However, after denoising, the gene variance in each stage of ageing becomes equal. (D) Here, the impact of noise on the calculated differential centrality is seen. Here is a comparison of the estimated gene network degrees for elderly and young pancreatic beta cells. Denoised expression-based networks predict fewer non-zero differential degrees than unfiltered expression-based networks do. (E) Enriched panther pathway terms for the top 500 genes in old and young pancreatic beta cells that exhibited the significant reduction in variance after denoising. (Mishra et. al. (247))

epithelia, where a few viruses seem to have the most damaging effects. There are two main categories of alveolar epithelial cells, often known as pneumocytes. The permeability barrier function of the alveolar membrane is significantly influenced by the type

1 alveolar (AT1) epithelial cells for main gas exchange surface of the lung alveolus. The synthesis of surfactants is a critical function of type 2 alveolar cells (AT2), which are the ancestors of type 1 cells. Since AT2 cells, also known as pneumocytes type II, are a popular target for many viruses, it's critical to understand their regulation mechanisms, given the significance of ageing.

The scRNA-seq profiles of cells obtained from old and young mice lung cell (97) were filtered using our approach of denoising. Multiple cell types in the lungs of young and old mouse showed increased robustness in the predicted gene network after graph wavelet-based denoising (Fig. 2.5A). Increased consistency in inferred gene networks from data sets reported by two distinct groups was also a result of graph-wavelet based denoising (Fig. 2.5B). Despite being denoised individually, the gene networks predicted for old and young cells overlaped more, which suggests that true interactions are more likely to be anticipated using our method. Therefore, there was less possibility that noise would overpower any gene-network-based differences between old and young cells. We examined ageing based alterations in PageRank centrality (nodes i.e Genes). Studying the change in PageRank centrality, which measures the "popularity" of nodes, has the ability to reveal changes in the effect of genes. First, we used Enrichr to conduct a gene-set enrichment analysis and estimated the differential PageRank of genes in young and aged AT2 cells. Young AT2 cells contained enriched elements for integrin signalling, 5HT2 type receptor-mediated signalling, the H1 histamine receptor-mediated signalling pathway, VEGF, cytoskeleton control by Rho GTPase, and thyrotropin activating receptor signalling in the top 500 genes with higher PageRank (Fig. 2.5C). Since the expression of oxytocin and thyrotropin-activating hormone (TRH) receptors in AT2 cells was low, we disregarded the oxytocin and thyrotropin-activating hormone-receptor-mediated signalling pathways as an artefact. Additionally, the gene-set for the 5Ht2 type receptor-mediated signalling pathway included genes that were associated with the phrases "oxytocin receptor-mediated signalling" and "thyrotropin activating hormone-mediated signalling."

Most of the enriched pathways have evidence of activation in AT2 cells in the literature. The expression of several 5-HTR, including 5-Ht2, 5-Ht3, and 5-Ht4, as well as their function in calcium ion mobilisation, have been demonstrated in alveolar epithelial cells type II (AT2) cells. However, there were very few studies that demonstrated their differential importance in old and young cells. Similar to this, Chen et al (113) demon-

strated that adult rat alveolar AT2 cells in primary culture secreted less pulmonary surfactant in the presence of a histamine 1 receptor antagonist. In AT2 cells, the VEGF pathway is active, and it is well known that ageing affects VEGF-mediated angiogenesis in the lung. Additionally, it is known that VEGF-based angiogenesis declines with ageing. (114).

In older mouse AT2 cells, we also conducted gene-set enrichment analysis for genes with higher pageRank. The terms that appeared among the top 10 most enriched in both the Kimmel et al. and Angelids et al. data sets for the top 500 genes with higher pageRank in old AT2 cells were T cell activation, B cell activation, cholesterol biosynthesis and FGF signalling pathway, angiogenesis, and cytoskeletal regulation by Rho GTPase (Fig. 2.5D). In terms of the enrichment of pathway words for genes with greater pageRank in older AT2 cells, the findings from Kimmel et al. and Angelids et al. data-sets overlapped by 60 %. In general, our study revealed that inflammatory response genes were more significant in older AT2 cells. The relevance of genes involved in cholesterol production has increased along with the inflammatory response, which suggests that ageing affects the quality of pulmonary surfactants generated by AT2. Al Saedy et al. have shown that elevated cholesterol enhances surface activity deficiencies brought on by lung surfactant degradation (115).

Additionally, we used Enrichr to analyse the genes that were differently expressed in aged AT2 cells. The phrases cholesterol production, T cell and B cell activation pathways, angiogenesis, and inflammation mediated by chemokine and cytokine signalling recurred for genes up-regulated in elderly AT2 cells compared to young. RAS pathway, JAK/STAT signalling, and cytoskeletal signalling through Rho GTPase, however, did not show up as enriched keywords for genes up-regulated in senescent AT2 cells (Figure 2.6). However, it has been shown in the past that ageing alters the equilibrium of the pulmonary renin-angiotensin system (RAS), which is linked to aggravated inflammation and increased lung damage (116). It is well established that the JAK/STAT pathway contributes to the reduction in surfactant protein gene expression brought on by oxidative stress in AT2 cells (Park 2012). Overall, our findings reveal that although age-related changes in the expression of genes implicated in key pathways may not be statistically significant, these genes regulatory effect may be.

We further investigated significance of transcription factors change in ageing AT2

cells in order to get more understanding. We discovered numerous pertinent TFs in the top 500 genes in the aged AT2 cells with a higher PageRank. To create a more restrictive list, we only took into account the TFs that showed a non-zero value for the difference in degree between the gene networks of old and young AT2 cells. With the Kimmel at el. data set, we discovered 46 TFs with a change in PageRank and degree related to ageing for AT2 cells (Fig. 2.5E). The findings of the pathway enrichment were consistent with changes in the centrality (PageRank and degree) of the TFs with ageing. In AT2 cells, RAS signalling is known to stabilise some proteins, such as *Etv5*, which has a greater degree and PageRank in aged cells (118). Inflammation in the lung's alveolar cells is known to be regulated by another TF Jun (c-jun), which has a larger effect on aged AT2 cells (119). In aged AT2 cells, we also discovered that *Etv5* and *Jund* were co-expressed (Figure 2.7). It is known that *Jund* participates in cytokine-mediated inflammation, and its role tends to grow in aged AT2 cells. *Stat4* shown a greater degree of PageRank in the old AT2 among the TFs *Stats 1-4* that are involved in JAK/STAT signalling. Older AT2 cells also seem to be more affected by androgen receptor (Ar) (Fig. 2.5E). It has been shown that androgen receptors are expressed in AT2 cells, though there may be aging-related variations in how they regulate cells (120).

A comparable investigation of the interstitial macrophages (IMs) in the lungs was also carried out, and we discovered literature supporting the activation of enriched pathways. As both seem to have greater pro-inflammatory response pathways, such as T cell activation and JAK/STAT signalling, the gene-set enrichment output for significant genes in older IMs was somewhat comparable to the findings from AT2 cells. Ageing in IMs, in contrast to AT2 cells, seems to promote glycolysis and the pentose phosphate pathway. In the past, Viola et al. (121) showed that higher glycolysis and pentose phosphate pathway activity levels were implicated in the pro-inflammatory response in macrophages. According to our findings, the RAS pathway was not substantially enriched for genes that are more important in aged macrophages. These findings suggest that the pro-inflammatory mechanisms that become active with age may differ amongst various lung cell types.

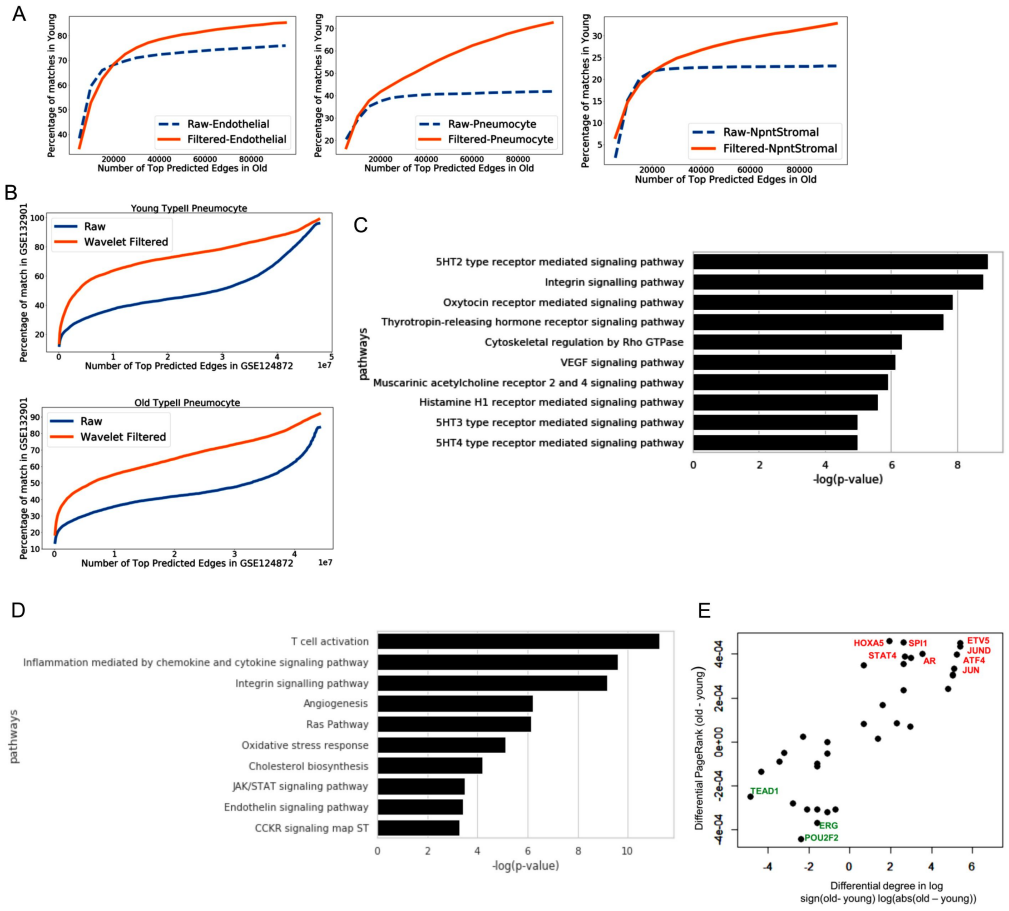

Figure 2.5: Improved regulatory inferences from single-cell transcriptome preprocessing of ageing lung cells using graph-wavelet (A) Reliability of networks predicted using the scRNA-seq profiles of young and elderly lung cells from Kimmel et al. . This figure show the coverage of the top 10,000 edges in young cells in the network inferred from old cells. After graph-wavelet based filtering, the estimated networks for old and young cells with the same type seem to have more overlap. The term "Raw" here denotes that the unfiltered scRNA-seq profiles were used to infer both networks (for old and young). Whereas the same result from the filtered scRNA-seq profile is shown. Utilizing correlation-based co-expression, networks were inferred. (B) The network overlap inferred from two distinct data sets, each with their unique batch effect, is shown. The X-axis displays the number of edges in the network inferred using the data set from Angelidis et al (GEO Id: GSE124872). The percentage of the top 10,000 edges in the network calculated using the Kimmel et al. data set is shown on the Y-axis. (C) The top 500 genes with the 10 most enriched Panther pathways have greater PageRank in young AT2 cells than in elderly AT2 cells. (D) The top 1000 genes with the 10 most enriched panther pathways had greater PageRank in older AT2 cells than in younger ones. (E) Scatter plot showing the difference in transcription factor (TF) PageRank (old-young) computed using networks predicted for old and young AT2 cells from the Kimmel et al. data set. Only TFs having a differential degree that is not zero are shown. (Mishra et. al. (247))

Comparison of the Impact of COVID-19 Infection and Ageing on Lung

An emerging pattern in the present SARS-COV-2 pandemic is that elderly people are more likely than younger people to have lung fibrosis and severity. We compared our findings to the expression profiles of lung infected with SARS-COV-2 reported by Blanco-Melo et al. (122) since our study demonstrated alterations in the effect of genes in lung cells owing to ageing. The host cell surface receptor for SARS-COV-2 attachment and infection, ACE2, has recently been shown to be primarily expressed by AT2 cells, according to (123). As a result, AT2 cells may be most significantly affected by COVID infection. We discovered that genes with considerable upregulation in the lung during SARS-COV-2 infection also had greater PageRank in the gene-network predicted for older AT2 cells (Fig.2.8A). We also used all varieties of lung cells collectively to perform the network inference procedure and calculate differential centrality between old and young. For genes that were up-regulated in the lung of SARS-COV-2 patients, we used gene-set enrichment. The majority of the 7 PANTHER pathway keywords were also enriched for genes with higher PageRank in aged lung cells in addition to being enriched for genes up-regulated in SARS-COV-2 infected lung (combined). In each of the two data sets utilised here, 6 out of 7 highly enriched panther pathways for genes up-regulated in COVID-19 infected lung were also enriched for genes with greater PageRank in older AT2 cells (5 in Angelids et al., 3 in Kimmel et al. data-based results).

Seven genes have a notable enrichment for genes with higher pageRank in aged AT2 cells among the top 10 enriched wikipathway keywords for genes up-regulated in COVID infected lung. However, in elderly AT2 cells, the term type-II interferon signalling did not significantly enrich for genes with greater PageRank. We also looked at the promoters of the genes that were up-regulated in the lungs of those with COVID infection. The top two enriched motifs for the promoters of genes up-regulated in COVID-infected lung belonged to the ETS and IRF families of transcription factors. It is known that *Etv5* is a constituent of the ETS group of TFs. Further investigation showed that COVID-infected lung has up-regulated expression of the majority of the genes whose expression is positively linked with *Etv5* in aged AT2 cells. In contrast, COVID-infected lung's mostly down-regulated genes that had a negative correlation with *Etv5* in aged AT2 cells. The *Stat4* gene had a same pattern. The pattern was the

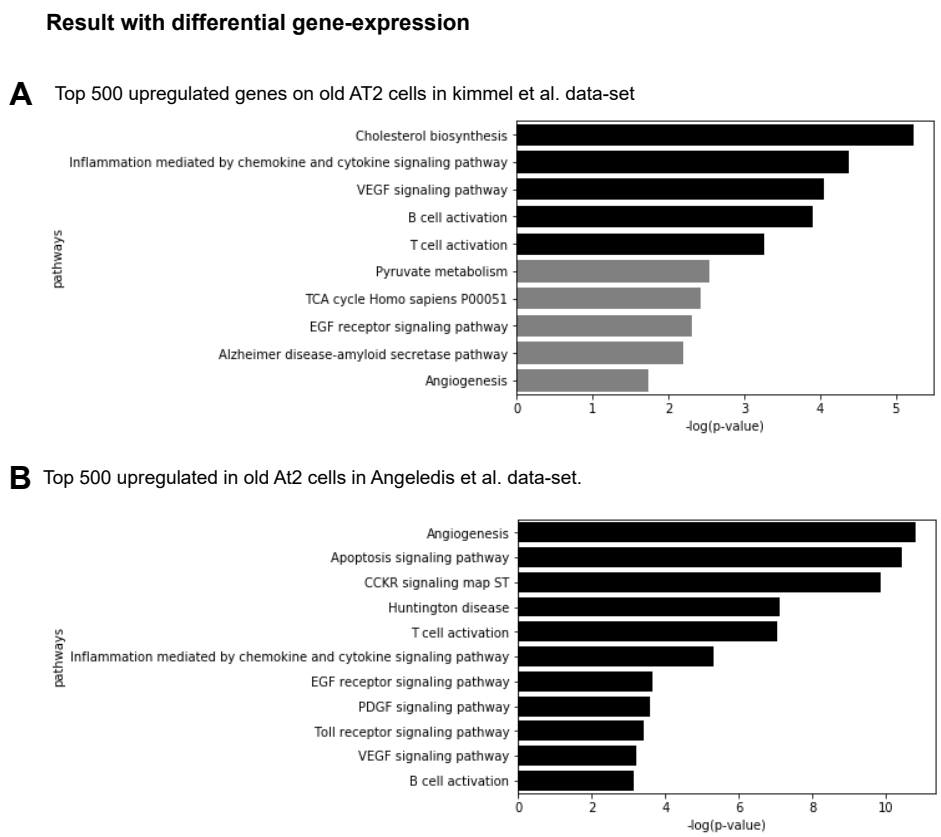

Figure 2.6: Panther pathway keywords were enriched for genes with substantially increased differential expression (FDR < 0.1) in aged AT2 cells vs young cells. The bar in grey indicates negligible enrichment (Pvalue < 0.05). Differential expression analysis results do not include some of the phrases that appear in PageRank-based results. RAS Pathway, JAK/STAT Signaling, and Cytoskeletal Regulation by Rho GTPase are examples of phrases that did not seem to be enriched for genes with greater expression in elderly AT2 cells. (Mishra et. al. (247))

inverse for the *Erg* gene, which had a higher pageRank in young AT2 cells. Positively linked genes with *Erg* in aged AT2 cells exhibited greater down regulation in COVID-infected lung compared to genes with negative association. This pattern suggests that a few TFs, such as *Etv5* and *Stat4*, with greater PageRank in aged AT2 cells, may play a role in the positioning or activation of genes that become more expressed by COVID infection.

2.4 Discussion

Single-cell expression profiles have a huge amount of potential for utilizing to infer regulatory changes in pure primary cells caused by ageing and other factors. These

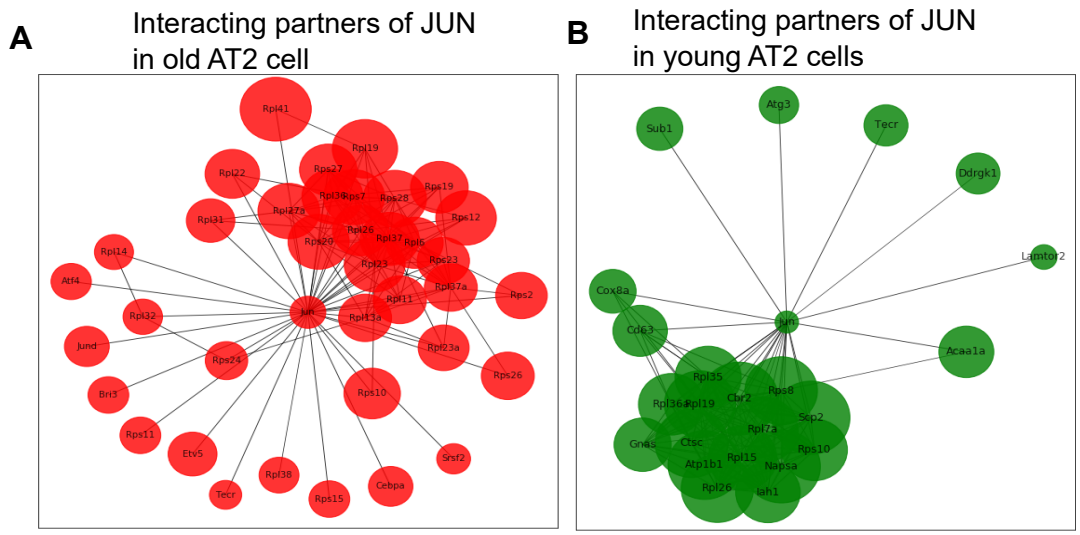

Figure 2.7: In both old and young mouse At2 cells, Jun genes interact with other genes. Ribosomal proteins make up the majority of Jun's interaction partners. Additionally, *Etv5* and *Jund* seem to be dependent on c-JUN via co-expression. (Mishra et. al. (247))

applications can include figuring out how a disorder develops or identifying signalling pathways and master regulators as possible therapeutic targets. Therefore, to aid biologists in their work-flow for graph-theory based analysis of single-cell transcriptome, we created GWNet to facilitate such investigations. By using a graph-wavelet based technique to decrease noise brought on by technological problems or cellular biochemical stochasticity, GWNet enhances the inference of regulatory interaction among genes. By comparing our filtering strategy to four benchmark data sets from the DREAM5 consortium and a number of single-cell expression profiles, we were able to show an improvement in gene-network inference. We demonstrated how our strategy for filtering gene-expression may aid gene-network inference approaches using 5 distinct methods for estimating networks. Our comparisons of graph-wavelet with alternative imputation, smoothing, and filtering techniques revealed that graph-wavelet is more adaptable to changes in the expression level of genes with shifting cellular surroundings. Thus, a theoretically distinct method for pre-processing gene-expression profiles is graph-wavelet based denoising.

There is a good amount of literature on using gene-networks inferred from bulk gene-expression profiles to compare and measure differences in two groups of samples. However, using conventional techniques on single-cell transcriptome profiles hasn't given significant results. Even though each data set was filtered individually,

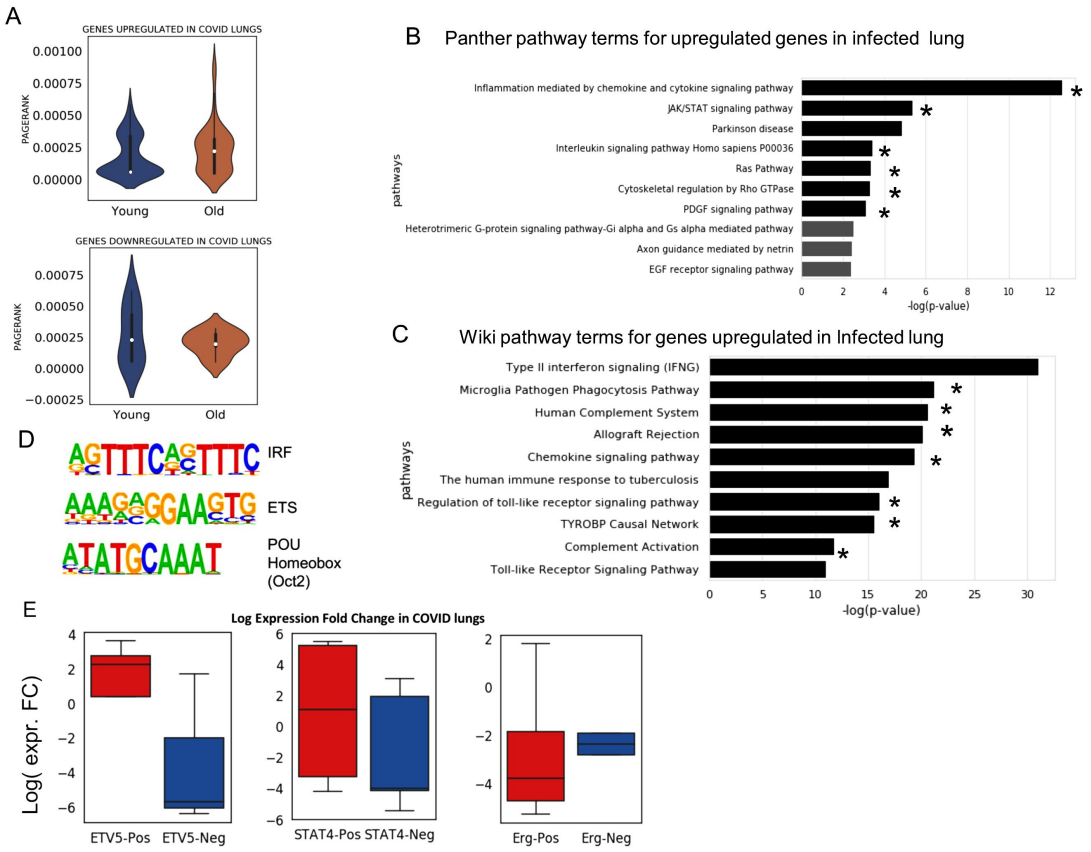

Figure 2.8: Analysis of scRNAseq profile in SARS-COV-2-infected lungs (COVID). (A) PageRank distribution of genes up-regulated in COVID-infected lung ($FDR < 0.05$) (122). For networks that were computed using scRNA-seq of both young and elderly AT2 cells, the PageRank is shown. For genes that were downregulated ($FDR0.05$) in COVID-infected lung, PageRank is shown in a similar fashion.(B) Top 10 Panther pathway genes that are abundant in COVID-infected lung. In older AT2 cells, the phrases indicated by an asterisk (*) likewise significantly enrich for higher pagerRank genes. (C) Top 10 wiki pathway for genes that have higher expression in COVID-infected lungs. In older AT2 cells, the phrases indicated by an asterisk (*) are likewise enriched ($Pvalue0.05$) for genes with higher pageRank. (D) The top 3 known transcription factors (TF) motif enrichments in gene promoters in COVID-infected lung. (E) Fold change in expression of genes with positive and negative correlations to transcription factors in aged AT2 cells in lung with COVID infection. Two transcription factors (TFs) *Etv5* and *Stat4*, which have increased PageRank in elderly AT2 cells, are shown. The results for *Erg*, which had higher PageRank in young AT2 cells, are also shown as a control. The majority of the genes whose expression correlated positively with those of *Etv5* and *Stat4* in aged AT2 cells were also up-regulated in COVID-infected lung. *Erg*, meanwhile, perceives the reverse trend. In aged AT2, genes with positive association to *Erg* genes were more down-regulated than those with negative correlation. Such findings imply a potential role for transcription factors (TFs) whose impact (PageRank) increases with age in either activating or posing the genes up-regulated in COVID infection. (Mishra et. al. (247))

our strategy seems to address this problem by boosting consistency and overlap across gene-networks inferred using an expression from diverse sources (batches) for the same cell type. After graph-wavelet based denoising of expression profiles, the inferred network from independently processed data sets from various sources shows an increase in overlap, suggesting that the estimated interactions among genes are approaching accurate values. It is more reliable to compare a regulatory pattern between two groups of cells when network prediction is closer to actual values.

Furthermore, Chow and Chen (124) recently shown that SARS-COV-2 infection is a significant inducer or suppressor of the expression of age-associated genes discovered utilising bulk lung expression profiles. They did not, however, analyse data using a systems-level perspective. Our results showed that genes with stronger influence on ageing AT2 cells and genes up-regulated in COVID-infected lung are enriched for the RAS and JAK/STAT pathways. RAS signalling is thought to be crucial for AT2 cell self-renewal (118). Similar to this, Yew et al (125) notes that the JAK/STAT pathway is known to be activated in the lung following damage and affect surfactant quality (214).

When we employed our method on murine aging-lung scRNA-seq profiles, our analysis offered a crucial insight i.e. elderly AT2 cells may be predisposed to have more RAS and JAK/STAT signalling because of their regulatory patterns and master-regulators. The androgen receptor (AR), which has been linked to male pattern baldness and a greater risk of men contracting COVID, had a higher pageRank and degree in elderly AT2 cells (126). However, further research is required to link the severity of COVID infection brought on by ageing and AR. Contrarily, we see a strong effect of genes implicated in Histamine H1 receptor-mediated signalling in young AT2 cells, which is known to control allergy responses in the lungs (127). Our kind of analysis also has the advantage of highlighting a few particular treatment candidates for future research. In therapeutic trials for pulmonary fibrosis, a kinase named JNK that binds to and phosphorylates c-Jun (128) is being examined. According to Montepoli et. al (129), androgen restriction treatment has been found to provide a limited degree of protection against SARS-COV-2 infection. In keeping with this pattern, our study suggests that *Etv5* may potentially be a potential therapeutic target to lessen the impact of age-related RAS pathway activation in the lung.

However, the method suffers some limitations as well. The search for an optimal

threshold for denoising linear signals and images using wavelet coefficients has been an active area of research. In our study, we examined both soft and hard thresholding approaches and investigated the minimum description length principle (MDL) which is an information-theoretic criterion. In this work, we predominantly utilized hard thresholding for most of the datasets. Soft thresholding of Graph-wavelet coefficients requires further refinement and is still an ongoing research topic. Additionally, for determining the hard threshold value, we explored the best overlap between predicted gene-network and protein-protein interaction (PPI) data. This approach offers a robust way to establish the threshold while identifying the frequency cutoff. However, this approach can be automated for finding optimal threshold which works best for eahc signal individually.

Secondly, validating the top predicted gene-gene interactions presents challenges due to the wide variation of regulatory mechanisms across different cell types or tissues. Consequently, establishing a ground truth for each set of predicted interactions becomes laborious and complex. Therefore, validating these interactions remains a difficult task, given the diverse nature of regulatory mechanisms across various biological contexts.

CHAPTER 3

Graph-based Integration of Large Single-cell Epigenome Profiles with Different Batch Effects

3.1 Background

Integration of several disparate modalities such as transcriptional and epigenetic profiles or integration across species or batch effect, is a crucial topic that poses considerable technical challenges. To integrate an ATACseq-based panel of measurements, several graph-based algorithms, including CONOS (clustering on network of samples) (152), employ a unified graph representation approach. These methods subsequently demonstrate the integration between ATAC-seq and RNA-seq datasets. However, they demonstrate that, despite the fact that such integration may be highly successful, it relies majorly on the data's resolution and the capability to establish a meaningful relationship between gene expression and other modalities.

Other reported methods that extract latent properties using generative frameworks based on deep learning or canonical correlation are unable to capture cell-type specificity across batches and species. A search engine methodology, on the other hand, may assist in processing each individual cell independently of the others, which has the huge advantage of retaining each cell's unique information and simultaneously making use of datasets from other research. Single-cell profiles may be searched using techniques like scfind (145) to find housekeeping and cell-type specific genes. A few other methods (146), (147) (148) (149) have combined single-cell epigenome with single-cell expression profiles, but they did not use the strategy of looking through a large pool of reference cells. Seurat (150), LIGER (151), and Conos (152) are other tools that have been suggested for integrating single-cell open-chromatin profiles.

The majority of integrative techniques, according to recent benchmarking research by (153), performed poorly for batch correction when integration of scATACseq profiles was performed. The study assessed more than 30 such single-cell integration methods.

Only 27 % of Leucken et al. integration's outputs for scATACseq profiles outperformed the best-unintegrated findings, revealing a peculiarity about such integrative approaches (153). Due to this, there is a lack of a reliable mapping method that can enhance the development of search engines intended to accurately match query single-cell epigenome profiles to a large number of single-cell profiles regardless of batch effect, despite the availability of large single-cell epigenome atlases (148) (149) (150) and methods for their low-dimensional visualisation.

We develop a Graph based methodology for joint embedding of cells from various batches, species, and peak-lists that includes a novel approach for computing the distance between cells by projecting them onto a common reference pool of dataset. The proposed technique resolves the issue of employing reference cell atlases for extremely effective graph-based joint-embedding of query single-cell open-chromatin profiles regardless of batch effect, species, and peak-list. This aids in eliminating the batch effect and learning a reliable neighbourhood of the cells using knowledge-based learning of edge weights among cells. The approach also offers an interactive visualisation of the joint embedding's KNN-based graph. ScEpiSearch also incorporates computational techniques for matching a substantial pool of reference single-cell expression profiles with the queried single-cell open-chromatin profiles. scEpiSearch is also able to handle non-identical peak-lists of single-cell epigenome profiles from diverse research groups and determine the statistical significance of the query's match with single-cell expression profiles. Instead of using gene activity (154) as a proxy for cell-type specificity, scEpiSearch calculates gene-enrichment score. In order to get a deeper understanding of regulatory behaviours via their epigenomes in pathogenic profiles, we also use scEpiSearch on single cell K562 cells and capture heterogeneity, lineage bias, and stress-response across single-cells.

Additionally, single-cell epigenome profiling allows to recognize poised and active cis-regulatory sites as well as the fundamental mechanisms governing genome regulation in a variety of in vivo and in vitro cell types and tissues. Single-cell epigenome profiling is increasingly being adapted for large scale atlas datasets and provides accurate insights into underlying cell state regulation due to its various advantages like lack of RNA degradation, digital readout of chromatin status and a better understanding of heterogeneity in cellular responses (155) (156). So handling the complexities of searching for and meta-analyzing single-cell epigenome profiles is crucial.

Such tasks can be handled by a search engine, which can also show distinct phases of cancer cell's de-differentiation and anticipate a cell's behaviour in an unknown condition. As the approach makes use of the additional information found in the reference pool of cells in various cellular states, such an approach has the ability to result in improved annotation and regulatory inference from query single-cell open-chromatin profiles. Lahnemann et al. (157) identified mapping a single-cell to reference atlas as one of the eleven great challenges in single-cell data science and emphasised the potential and pitfalls of such techniques. Another major problem they identified is combining single-cell data from many samples and trials. Several (145) groups (146) (147) have made an attempt to develop a search engine for single-cell expression profiles. They do not, however, address the problems posed by similar obstacles for single-cell epigenome datasets.

Single-cell open-chromatin profiles provide unique challenges compared to the single-cell transcriptome. Currently, single-cell epigenome profiling primarily aims to capture open chromatin regions(preferentially at promoters, enhancers, etc.) using MNase-seq (Micrococcal Nuclease digestion with deep sequencing) (159), ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing), DNase-seq (DNase I hypersensitive site sequencing) (158), or ATAC-seq (154) When compared to an equivalent matrix for a single-cell expression dataset, the read-count matrices constructed using single-cell open-chromatin contain more genomic loci (peaks) as features. The read-count matrices of single-cell open chromatin profiles generated by several research teams often include genomic locations (peaks) that are distinct from one another. Therefore, single-cell open-chromatin profiles cannot be directly analysed using the current algorithms and search techniques established for single-cell expression profiles.

Hence, in our second contribution, we introduced a unique graph-based embedding method of scATAC-seq and scRNAseq. This method incorporates a method which computes similarity among cells and provides a graph based visualization enabling researchers to gain insights into the underlying biological processes. It also handles single-cell ATAC-seq profiles from diverse sources, irrespective of the species of origin or batch effects. By overcoming species and batch effects, our method enables the investigation of diverse datasets and facilitates the discovery of important regulatory patterns and cellular behaviors. This contribution contributes to advancing our understanding of complex biological systems and has potential implications for the development of

targeted therapies and treatments in cancer research.

3.2 Material and Methods

3.2.1 Pre-processing Reference Profiles

Using the same set of genes, datasets from several studies on human/mouse cells are integrated. Similar to this, single-cell expression profiles from mouse samples were convened together. The quantile-normalized of single-cell RNA-seq based expression (FPKM) profiles were performed as pre-processing step(160). The mean expression of all reference cells for each gene is then computed. The expression of a gene is normalised for each reference cell by its average expression across all cells. Thus, we change the relative expression of genes for each cell to their cell-specific expression.

Instead of using any dimension reduction or latent representation directly, we employed cell-type annotation to cluster single-cell expression data. Cell-type annotations were previously available for a number of human and mouse scRNA-seq datasets. We conducted annotation for scRNA-seq data without cell-type information using matchscore2 (165). The KNN-based approach was used to identify sub-clusters of cells in each main cluster of cells created using cell-type annotation. As their representative, the vectors expressing the average scores of gene expression in cells that form a subcluster were chosen. We created nearly 30000 representative expression vectors for 3 million or more human and mouse scRNA-seq profiles. Hence, the representative mean vectors became less sparse as a result of this technique, and it also sped up the hierarchical search for the expression profile that matches a given query. We used Python in both the standalone and web server implementations of scEpiSearch. The scEpiSearch architecture is flexible enough to accommodate extremely large reference single cell profiles. Approximately 3.3 million cells altogether, including 1288653 human single cells and 2141797 mouse single cells (scRNAseq), make up the current version of scEpiSearch.

3.2.2 Single-cell ATACseq Pre-processing to be Projected on Reference RNA-seq Profiles

Initially, the query is pre-processed by dividing each peak's scATACseq read count by its global accessibility score, thus, revealing cell type-specific peaks (mainly enhancers) in query cells (equation (3.1)).

$$t_{ij} = r_{i,j}/(a_i + \epsilon) \tag{3.1}$$

The r_{ij} is the read count with i^{th} peak in j^{th} single cell which is normalized by its global accessibility score a_i with pseudo-count ϵ added to it.

It identifies genes nearest to peaks in the scATACseq query profile. It initially makes advantage of the pre-existing table of genes proximal to the global accessibility peak list in order to efficiently identify proximal genes. Every query scATACseq profile has overlaps between its peaks and sites on the global accessibility list, according to scEpiSearch. Our evaluation also helped determine that for 65 to 80 percent of query peaks, scEpiSearch successfully achieved overlap with the global accessibility peak list. ScEpisearch optimizes the proximal genes from pre-existing tables for peaks that coincide with regions in the global accessibility list. If accurate mode is selected, then it separately searches for proximal genes individually for peaks that did not have an overlap with the global accessibility peak list. As a result, for the majority of the peaks, scEpiSearch does not repeat the process of seeking proximal genes.

scEpiSearch calculates gene enrichment (GE) scores for each query cell by selecting genes proximal to peaks with higher normalised read-count as foreground and leaving genes nearby to all other peaks as background. Fischer's exact test determines gene-enrichment scores for each query cell using foreground and background genes, as shown in equation (3.2).

$$\sum_{i=k_m}^{min(n,K_m)} \frac{\left(\frac{K_m}{i}\right)\left(\frac{N-K_m}{n-i}\right)}{\left(\frac{N}{n}\right)} \tag{3.2}$$

Here, K_m is the frequency with which a gene m occurs in the background set, and N denotes the total number of times that all genes in the background have been observed. The number of times the gene m occurs in the foreground is km. The total

number of all foreground genes is n at the same moment. As a result, the p-values of the genes around numerous cell type-specific sites—mostly enhancers—are lower (and hence more substantial). It is to be noted that this is the gene-enrichment calculation and that it is distinct from the gene-set enrichment carried out by other tools (162) (163)

3.2.3 Finding Matches for Query Cells in Reference Expression Profiles

The methodology for searching the scATACseq profile for a similar expression profile is based on the known evidence that genes close to enhancers have significant cell-type-specific expression. To highlight genes commonly detected proximal to putative enhancers or cell-type-specific loci, scEpiSearch first calculates gene enrichment scores of the query cells. Using representative expression vectors for clusters of scRNA-seq cells, ScEpiSearch calculates MExTEG (Median expression of top enriched genes) so that for query cell q the MExTEG value of query in reference cell m is

$$MExTEG_q(m) = median(Expr_m(top\ enriched\ genes\ in\ q)) \tag{3.3}$$

Furthermore, it should be emphasized that we employ cell-type-specific expression rather than the raw transcripts (FPKM or TPM) of cells or the representative vectors of cells. By dividing a gene's expression in a cell by the median expression score of the same gene across all cells, the cell-type specific expression of a gene is obtained. Therefore, MExTEG reflects the median score of cell-type-specific expression of the top enriched genes for query scATACseq profile for a reference scRNA-seq profile. The MExTEG value is initially computed using an expression vector for the cluster of reference scRNA-seq profiles due to the hierarchical nature of the search for matching expression profiles. Another point to be noted is that for each cluster, the representative expression vector comprises the mean of the cell type specific gene expression profiles for each cell in the cluster.

Moreover, using the null model of representative expression vectors, the MExTEG for query cells is transformed to a P-value. The top N clusters with the lowest MExTEG p-value are then selected. Additionally, MExTEG is once again calculated for

the query cells using a cell-type-specific expression profile of single-cells found in the top N chosen clusters. The top 100 reference cells with highest MExTEG values are selected. Using a null model, the MExTEG is transformed to a p-value score for 100 reference cells with higher MExTEG for the query. The null model described below is used to determine the p-value for these matches. A new P-value is computed as the final refinement score based on rank computed using P-value-ranks. Below are the specifics of how two different P-values were calculated.

Statistical Approach to Calculate the Significance of Match of Query in Reference

We used a few randomly chosen normalised scATACseq profiles (normalized by global accessibility scores of peaks) to create a null model for our search process. The top 1000 genes with the highest gene-enrichment score were retrieved for the 500 cells that were randomly chosen. The highest enriched gene lists of two cells in a pair were combined after 500 randomly chosen cells were divided into pairs. 1000 genes were chosen at random from the combined list of 2000 genes for each pair of cells. As a result, we created thousand query vectors with 1,000 genes each that were used as false query cells. Using the reference cell's cell-type specific expression profiles, $MExTEG$ is computed for each cell in the set of false queries to produce a matrix with the dimensions $1000 X No of reference single-cells$. We determine the p-value of similarity for a reference cell as the proportion of null model cells (false queries) that have greater values of $MExTEG$ than the query cell. Consequently, the P-value of the match between the query q and the reference expression profile of cell m is determined as follows.

$$Pval_q(m) = \frac{Number\ of\ Null\ model\ cells\ with\ MExTEG(m)\ >\ MExTEG_q(m)}{1000} \tag{3.4}$$

Rank-based statistical method to improve match significance: The rank of matches owing to P-values computed by MExTEG is further adjusted or refined using scEpiSearch in order to decrease bias in the search for matching expression cells. We retain the pre-calculated rank of each reference cell for all cells in the null model for this reason (false queries explained above). A perspective of bias in the data is provided by this rank computation, which also counts the instances in which a reference cell is a top hit for cells in the null model. As a result, we determine a new P-value after computing the

match's P-value and ranking the reference cell for the query cell. The proportion of cells in the null model for which the identical reference cell has a better rank than the query cell is used to determine the new P-value of the match between a reference cell and a query cell.

3.2.4 Finding a Match in Mouse Reference Expression Profiles for Better Graph Learning

Additionally, the mouse reference scRNA-seq dataset and the human cell scATACseq profile may be matched using scEpiSearch. It is based on the idea that because the same markers are often used to distinguish distinct cell types in both mice and humans, it is extremely probable that the cell-type-specific expression of genes in the same cell type from two species would be comparable. Because of this, scEpiSearch employs a strategy that emphasises enriched genes close to foreground peaks (potentially enhancers) with a high degree of cell-type specificity.

Therefore, it is theoretically conceivable to query the scATACseq profiles of human cells and discover the appropriate matching mouse reference expression profiles using our technique. In order to query the scATACseq profile of human cells, scEpSearch translates read counts to gene enrichment scores. It computes MExTEG utilising the reference single-cell expression profile of mouse cells for the human cell query. P-value is obtained using a precalculated $MExTEG$ value for a null model created using human scATACseq profiles and a mouse cell reference expression dataset. scEpiSearch creates a new P-value for the match with a reference mouse expression vector based on its precalculated rankings for the null model created from human cells after determining the rank of reference mouse cells for the query human scATACseq profile based on MExTEG based p-value.

3.2.5 Graph Based Embedding of Multiple Query Single-cell ATACseq Profile

In order to manage input multiple query read-count matrices with various peak-lists from both human and mouse cells and achieve their co-embedding, we developed a

distinct approach that is incorporated in scEpiSearch. For this reason, scEpiSearch independently calculates gene enrichment scores for each batch of the query scATACseq profiles with various peaks/species. Different scATACseq read counts are integrated into the same feature set since enrichment scores are computed for all genes in the same list. As a result, scEpiSearch combines GE scores from several read-count matrices into a single matrix. The batch effect has an impact on scEpiSearch when GE scores are directly used, hence it does not directly employ GE for embedding. The method outlined above is used to find matches for queries in mouse single-cell expression profiles after GE scores from several read-count matrices are combined.

The mouse null model is used to mouse query cells. Similar to how cross-species search was described above, scEpiSearch employs a null model created using human cells for human query cells.

Calculation of Distance between query cells: For smart Graph-based embedding of queries, scEpiSearch creates a network with each node representing a query cell after identifying matched mouse cells for each and every query cell. It creates an edge with a weight equal to the number of top matching reference cells in the same cluster that links two query cells (nodes). Take the query cell X as an example, where 4 of the top 10 matched mouse expression profiles are members of the sub-cluster A. The edge connecting query cells X and Y would have a weight of 4 if another query cell Y has 5 out of the top 10 matching expression profiles from the same sub-cluster.

Graph Visualization of Query cells Embedding: The associations of a node with just K neighbours are retained after computing all edge weights. In order to get the 2D coordinates of nodes, Fruchterman and Reingold technique is used after creating a KNN-based network with edge weights. Force-directed graph drawing is accomplished using the Fruchterman and Reingold algorithm (164). When creating a KNN-based graph, we choose an appropriate value for K (number of neighbours) in order to use the Fruchterman and Reingold technique to prevent overlap between different cells. The 'networkX' library ($draw\ network\ nodes\ function$) is used to visualise the network after calculating the 2D coordinate of the nodes (representing query cells) (165). It further conducts spectral clustering (166) to aid users by locating cell clusters using a KNN-based network of queries.

Fruchterman and Reingold algorithm: The algorithm considers two principles

as basis for the node positioning i.e 1) Vertices connected by an edge should be drawn near each other. and 2) Vertices should not be drawn too close to each other.

It calculates k, the optimal distance between vertices as

$$k = C\sqrt{\frac{area}{No\,of\ vertices}} \tag{3.5}$$

where C is constant.

Spectral Clustering to assign clusters to cells in embedded graph The foundations of spectral clustering may be found in graph theory, which uses the method to find communities of nodes in a network based on the connections that connect them. The information used in spectral clustering comes from the eigenvalues (spectrum) of unique matrices created from the graph or data collection.

Spectral clustering employs clustering on a projection of the normalised Laplacian. When a cluster's centre and spread cannot adequately describe the whole cluster, as is the case when clusters are nested circles on a 2D plane, or more broadly when a cluster's structure is substantially non-convex, spectral clustering is particularly helpful in practice.

This technique may be used to discover normalised graph cuts if the affinity matrix is the graph's adjacency matrix (246). A kernel function like the Gaussian (also known as RBF) kernel with Euclidean distance is used to generate an affinity matrix, or the user may give their own pre-computed affinity matrix. We provide our own created affinity matrix to 'sklearn' function.

Calculation of Clustering Purity

We used 'DBSCAN' to cluster the 2D coordinates of the cells in order to compare the 2D embedding from various approaches (Figure 3.4). For instance, n clusters would be 2 if human and mouse B cell and T cell types were present. To assess if the embedding was done correctly, we employed two metrics. The adjusted Rand index and Normalized mutual information are the metrics used for evaluation.

Adjusted Rand index (ARI) : Let, $T = [t + 1, ..., t_P]$ represents the true p classes consisting of no of observations as n_i in class t_i and $V = [v_1, ..., v_K]$ are the clustering

result with k clusters having n_j number of observations in cluster v_j. ARI is determined by:

$$\frac{\sum_{i=1}^{p}\sum_{j=1}^{k}\left(\frac{n_{ij}}{2}\right)-\left[\sum_{j=1}^{p}\left(n_i 2\right)\sum_{j=1}^{k}\left(n_j 2\right)\right]/\left(\frac{n}{2}\right)}{\left(\frac{1}{2}\left[\sum_{j=1}^{p}\left(n_i 2\right)\right)+\sum_{j=1}^{k}\left(n_j 2\right)\right]-\left[\sum_{j=1}^{p}\left(n_i 2\right)\sum_{j=1}^{k}\left(n_j 2\right)\right]/\left(\frac{n}{2}\right)} \tag{3.6}$$

where,

$$n=\sum_{j=1}^{k} n_j=\sum_{i=1}^{p} n_i \tag{3.7}$$

Normalized mutual information (NMI): Normalized Mutual Information (NMI), the second metric we used, is computed as

$$\frac{2I(U,V)}{H(U)+H(V)} \tag{3.8}$$

Here, $I(U,V)$ is mutual information and $H(U)$ and $H(V)$ are the entropies of U and V are cluster labels. We utilised the adjusted_rand_score() and normalized_ mutual_info_score() methods from 'sklearn' in the python programming language to calculate the ARI and NMI.

Comparison of Embedding with Other Tools

Other techniques for scATACseq profile analysis do not look through reference datasets to seek comparable cells. Few researchers have, however, suggested utilising their techniques to integrate several scATACseq profiles from diverse sources into a single visualisation display. We downloaded these tools and evaluated how well they performed when we tried to incorporate scATACseq profiles from various sources. The sources they utilised and the criteria they used are described below.

- SCALE: The URL for SCALE (190) is [1]. The following settings were used to execute it: $batchsize$ = 500, $seed$ = 43, $minpeaks$ = 400, lr = 0.0002, and $maxiter$ = 500. We merged the results of each query's latent space representation and then submitted them to tSNE to get the final visualization.
- SCVI: The URL for SCVI (189) is [2]. The model learnt the latent representation of the query cells after receiving the pooled gene enrichment scores of all queries

[1] `https://github.com/jsxlei/SCALE`
[2] `https://github.com/YosefLab/scvi-tools`

which we generated. To get a scatter graph, we further passed these latent attributes to tSNE.

- SCANORAMA: The URL for SCANORAMA (187) is [3]. In order to get integrated form for cells for SCANORAMA as well, we aggregated the gene enrichment scores of all queries supplied to the $scanorama.correct()$ method.
- MINT: By installing the mixOmics library in R, accessible at [4], MINT (188) was used. Gene enrichment scores of the query cells were supplied to the function $mint.plsda()$ in MINT, and their function plotIndiv was used to create the plot.

3.2.6 Evaluation of Co-Embedding of Single-cell ATACseq and RNAseq Profiles Across Species

Despite the fact that integrative techniques for scATACseq yielded to unsatisfactory findings reported by Leucken et al. reported (171). We compared our method with three distinct approaches (Seurat, LIGER, and Conos) since we wanted to be certain and compare them to scEpiSearch. The parameters utilised by various techniques are listed below.

- Seurat: Details on Seurat 3.2 (150) and the ATACseq integration vignette.html can be found at [5]. The Seurat object was constructed after the RNA-seq reference count data and ATAC-seq queries were loaded into R. On the RNA-seq data, further conventional analyses were carried out (normalization, finding variable genes, scaling data and running PCA, tSNE). Separate ATAC-seq data analysis is done, and gene annotation data is added to the Seurat object. There were anchors determined between the two modalities. Canonical correlation analysis uses the gene activity scores and scRNA-seq gene expression quantifications with all the genes that were highly variable in the RNA-seq dataset (parameters being reduction = "cca", k.anchor = 5, k.filter = 80). Utilizing the option $weight.reduction = human.atac[[“lsi”]]$, the label transfer was accomplished. Finally, using pre-calculated anchors, RNA-seq is imputed into the scATACseq for visualisation, and the datasets are then combined with parameters using the default settings. It is to be noted that human version of the single-cell ATACseq read-count matrix was in hg19 format while the mouse version used the mm9 format.
- LIGER: Linked Inference of Genomic Experimental Relationships was used with details available at [6] ATAC-seq data was transformed into gene counts so that

[3] https://github.com/brianhie/scanorama

[4] http://www.bioconductor.org/packages/release/bioc/html/mixOmics.html

[5] satijalab.org/Seurat/archive/v3.2

[6] (https://github.com/welch-lab/liger/blob/master/vignettes/Integrating_scRNA_and_scATAC_data.html)

they could be compared to $RNA - seq$ which were obtained by counting total number of ATAC-seq reads within gene and promoter region (3 kb upstream) of each gene in each cell. Then after loading read counts in R, LIGER object was created with createLiger function. Both datasets are normalized using normalize function. Highly variable genes are identified and combined from both datasets. The parameter datasets.use was set to 2 such that genes could be selected from RNA-seq dataset in function selectGenes. Joint matrix factorization (iNMF) was performed on the normalized and scaled RNA and ATAC data using optimizeALS function with value of k being 20. Finally to fully integrate datasets quantile normalization was performed through $quantile_norm$ function with value of $knn_k = 5$. runUMAP function was used to get coordinates for each cell in integrated visualization with parameters being $distance = `cosine', n_neighbors = 30, min_dist = 0.3$. Notice that here we used the hg19 version of our read-count/peakfiles.

- Conos: The github repository located at [7] was used to install the Conos (152) R package. The tutorial's instructions for integrating RNA and ATAC-seq [8] were abided to while implementing. For ATAC-seq, the gene activity scores produced by Seurat v3.2 were utilised. The pagoda2 package was used for the basicP2proc preprocessing phase. With the following parameters: k=15, k.self=5, k.self.weigh=0.01, ncomps=30, and n.odgenes=5e3, the buildGraph function was also utilised. The $embedGraph()$ method was used to produce the embedding, and largeVis was used to get the coordinates.

- ScEpiSearch: Reference dataset was prepard as per the steps mentioned in methods section. With prepared reference mouse RNA-seq datasets, a cross-species search was conducted for ATAC-seq query datasets. The top three matches were chosen and filtered using a 0.05 p-value threshold. The average of the coordinates of the top three matches in the reference dataset was used to generate the coordinates for each query cell. R is then used to plot the combined visualisation of the reference and ATAC-seq queries. We utilised mouse cell atlas (MCA) dataset tSNE coordinates of reference cells, which were supplied by Chawla et al., for the majority of the figures. To ensure that our results could be reproduced by other researchers, we also conducted tSNE-based dimension reduction for reference cells. For example, we calculated new tSNE coordinates for the reference cells in Figure 3.3B.

Silhouette Coefficient Calculation

Another measure to evaluate clustering in order to calculate how close a data point is to its own anticipated cluster when compared to other clusters is measured by the silhouette index or coefficient. The mean intra-cluster distance (a) and the mean nearest-cluster distance (b) are used to determine the silhouette coefficient for each sample, as

[7] `https://github.com/kharchenkolab/conos#basics-of-using-conos`
[8] `http://pklab.med.harvard.edu/peterk/conos/atacrna/example.html`

$$S_i = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{3.9}$$

Where $b(i)$ is the mean distance between the i^{th} sample and all samples in the nearest cluster (other than the cluster of the i^{th} sample), and $a(i)$ is the average distance between the i^{th} sample and all other samples in the same cluster. We used the cluster package in R [9] to calculate the silhouette coefficient. For this, we considered queries and their related cell types from a reference in one class, and the remaining cell types were given their corresponding clusters or cell labels. For instance, if the reference and query cells were both macrophage cells, the reference and query cells were placed in the same cluster, while the other cell types were given their own cluster labels. We took into account the tSNE coordinates of the cells from the TSNE plot of each approach when determining the distance between cells.

3.2.7 Dataset Sources

The GEO database provides the scATAC-seq mouse dataset that was used to create the query sets (GSE111586). The GEO ids for the HL60 and K562 cell lines scATACseq profiles, which were utilised to examine the lineage, are GSE109828 and GSE65360, respectively.

We used the following dataset for embedding case studies: Human Neuron (GSE 97942), cells from Mouse Forebrain (GSE100033), Human HSC (GSE96769), Mouse HSC (GSE111586), Human Myoblast (GSE109828), Human GM12878 (GSE109828), Mouse B-cell (GSE111586), and Human GM12878 (GSE68103), Human Neuron (GSE 97942) (GSE111586)

We carried out the embedding of scATACseq profiles (Figure 3.8) of cells from two patients with multiple phenotype acute leukaemia (MPAL), peripheral blood mononuclear cells from healthy people (GEO id: GSE139369), progenitors of hematopoietic cells (GEO id: GSE96772), T cells (GEO id: GSE107817), and B cells (GEO (GEO id: GSE109828). We employed the scATACseq profile of progenitors of hematopoietic cells in blood, which included the progenitors MEP (megakaryocytic-erythroid),

[9] `https://www.rdocumentation.org/packages/cluster/versions/2.1.2/topics/silhouette)`

CMP (common myeloid), CLP (common lymphoid), GMP (granulocyte-monocyte), and MCP (mast cell progenitor).

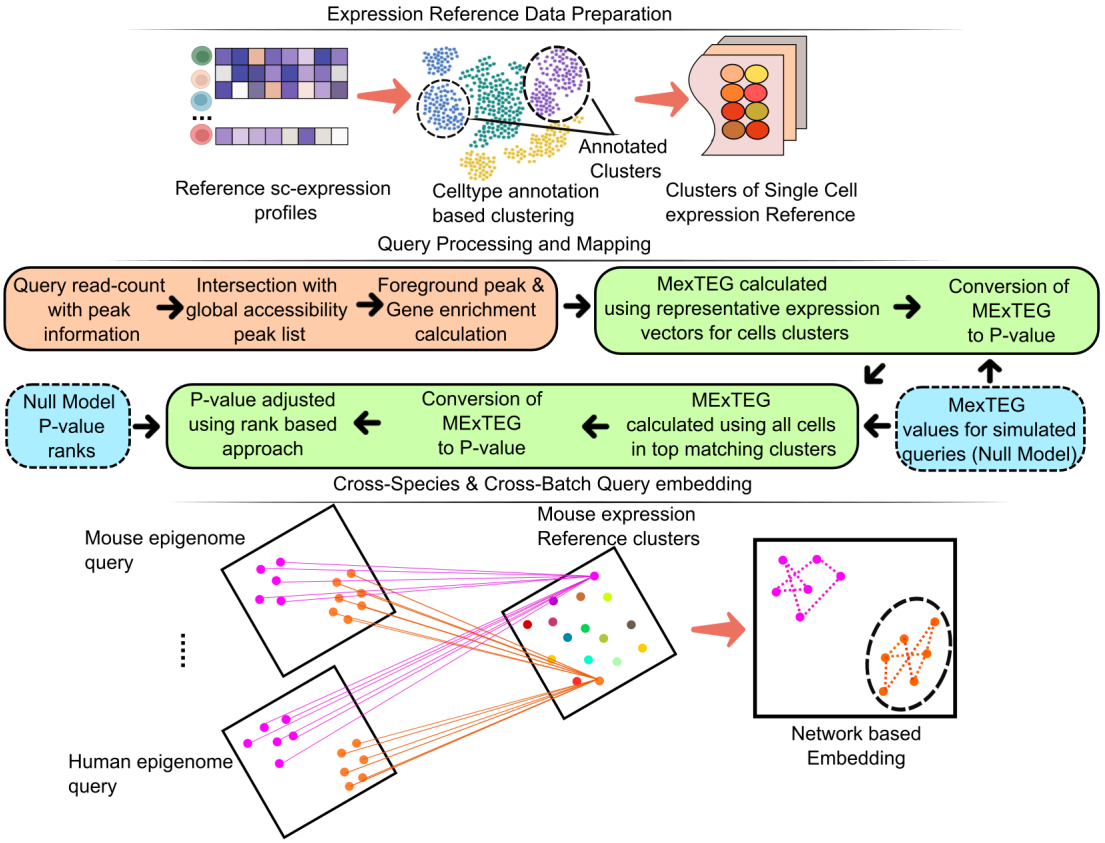

Figure 3.1: An illustration of the proposed method in scEpiSearch for improved regulatory state inference and annotation of query scATACseq utilising a large pool of single-cell epigenome data sets. It consists of the following steps: Expression Reference Data Preparation, Query Processing, Mapping and Projection Based Cross-Species and Cross-batch Query embedding. The cross-species and cross-batch query embedding provides co-embedding of various open-chromatin profiles utilising existing reference single-cell profiles, regardless of variation in peak-list in read-count matrix, batch effect, and species. (Mishra et. al. (248))

3.3 Results

The reference set of sc-expression profiles undergo initial preprocessing using scEpi Search. In practice, it also maintains a reference collection of single-cell transcriptomes and epigenomes that it has analysed (Figure 3.1). Nearly 3.3 million single-cell expression profiles from both human and mouse cell lines and over 8,00,000 cell lines make up the existing reference pool of scEpiSearch. The single-cell expression and epigenome profiles in the reference pool are handled by scEpiSearch in a clustered format to enable

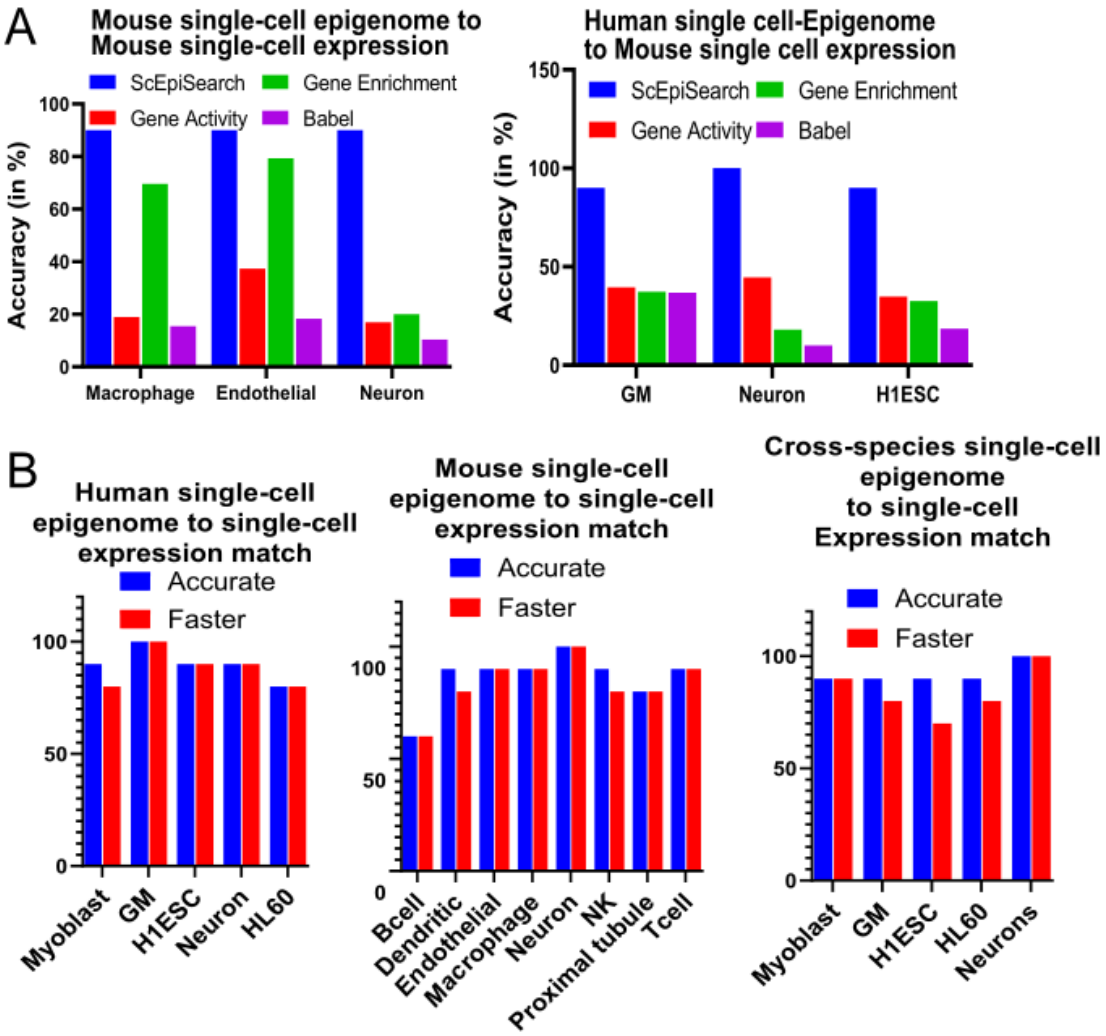

Figure 3.2: A) Evaluation of scEpisearch in comparison to 3 other methods based on the retrieved gene scores correlation for matching query scATACseq to a collection of reference single-cell scRNAseq. 10,000 MCA (mouse cell atlas) cells were selected as the scRNAseq reference dataset (MCA). Here, accuracy displays the proportion of query cells with the correct cell-type among the first 5 matches.B) Accuracy using the scEpiSearch for queries on the scATACseq read-count matrices is shown as follows, from left to right: i) query human scATACseq to reference human single-cell expression ii) search the mouse reference scRNAseq using the mouse scATACseq. iii) cross-species search – reference mouse scRNAseq using human scATACseq. The proportion of query cells for which the right annotation was one of the top 5 results is shown on the Y-axis.For the scEpiSearch modules for both faster and accurate, accuracy is shown as bar-plots. (Mishra et. al. (248))

hierarchical searching (Figure 3.1).

To identify possible enhancers while querying single-cell open-chromatin profiles, scEpiSearch initially normalises the read-count of each peak using a global accessibility score (172). We utilised the global accessibility peak-list created from numerous public

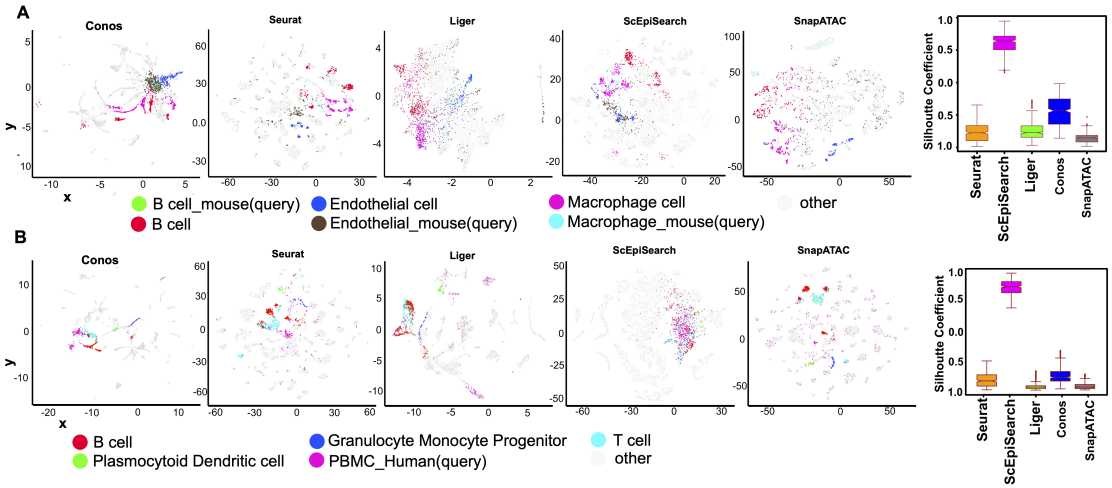

Figure 3.3: (A) A comparison of scEpiSearch with integrative approaches, using scRNA-seq profiles of 10,000 cells from the mouse cell atlas. Here, the search was restricted to the scATACseq profiles of three mouse cell types: B cells, macrophages, and endothelial cells. T he silhuotte index of the query cells for vicinity to the correct reference cell-types is shown. (B) Analysis of cross-species integrative approaches and scEpiSearch employing reference scRNAseq from MCA and human PBMC scATACseq profiles as the query. Also, human PBMC silhouette coefficients are shown. For the purposes of calculating the silhouette coefficients, immune cells and query cells were assumed to belong to single class in the references, whereas other cell types were regarded to be in other class. (Mishra et. al. (248))

open-chromatin profiles of bulk samples released by various groups and consortiums for both the human and mouse species (173) , (174). The bias is eliminated by normalisation by global accessibility score for peaks, which may have originated from different cells in the same query. As a result, each cell in the query is handled separately. The proximal genes are detected easily using a predetermined peak list in the fast version of scEpiSearch. The process of finding proximal genes is 1000 times faster for queries with more than 200000 peaks when intersecting with a predetermined peak-list with global accessibility, with almost 75-90 percent of peaks covered most of the time.

3.3.1 Reference Supported Improved Distance Calculation Despite Species Batch Effect

When scEpiSearch calculates the gene-enrichment score using the Fisher exact test, it takes peaks with high normalised counts as the foreground and others as the background set (hypergeometric test). It utilises its normalised expression values (explained in Methods) to determine the median expression for the top 1000 enriched genes (MEx-

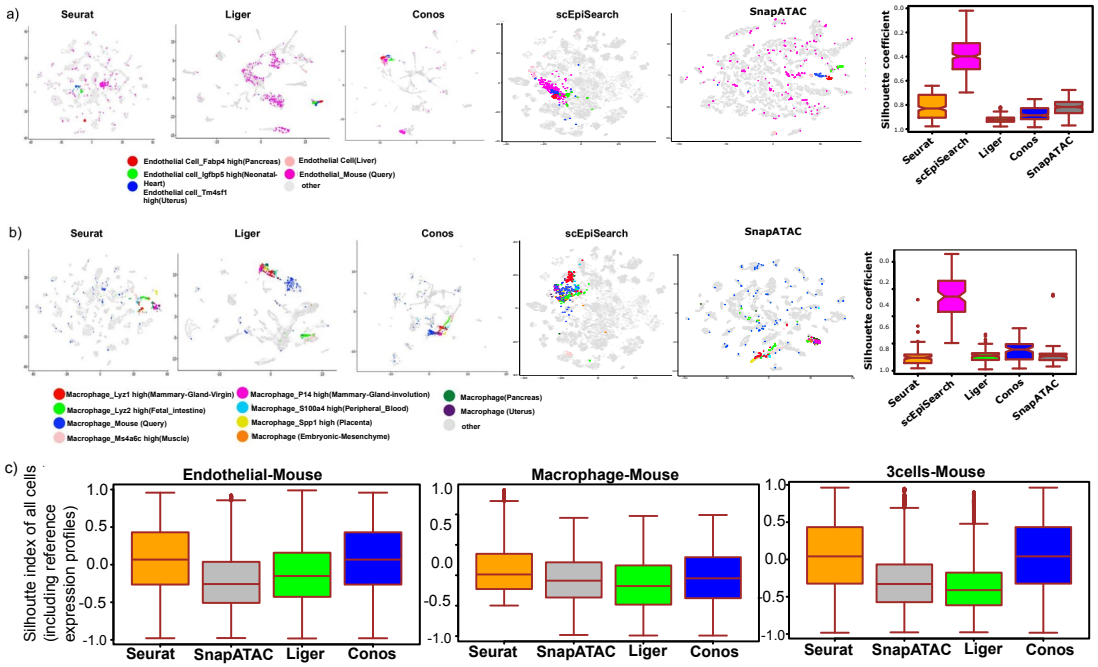

Figure 3.4: Utilizing reference expression and the query scATACseq from mouse cells, scEpiSearch is compared to integrative approaches. A) Mouse endothelial scATACseq profile was the query (Cusanovich et al.). For each of the 4 approaches, the query cells' silhouette coefficients are shown. B) Mouse macrophage epigenome profiles query (Cusanovich et al.) are used. Also shown are the silhouette coefficients for the query cells. C) Incorrect assessment of the 2D embedding figures when all cells, including the reference MCA cells silhouette coefficients are considered. Mouse endothelial cells, mouse macrophages, and 3-cell combinations are shown by the corresponding labels on the subfigures. (Mishra et. al. (248))

TEG) of query cells in order to match to a reference single-cell transcriptome profile. Using precalculated MExTEG values for cells in the null model, the MExTEG of a reference cell is translated to a P-value for the query (Figure 3.1). The scEpiSearch algorithm employs a hierarchical method to first match the query with representative expression vectors for clusters of cells instead of comparing to a vast pool of reference single-cell transcriptome profiles. The MExTEG values and accompanying P-values are then calculated using reference cells from the top matched clusters. To lessen bias in the dataset and search methods, scEpiSearch also computes a new P-value based on the rankings of reference cells for a query (Figure 3.1). In other words, scEpiSearch uses its pre-calculated rankings for the null model to change the rank of results.

Utilizing a reference set of 10,000 mouse single-cell expression profiles from the mouse cell atlas (MCA), we first evaluated our approach. We contrasted it with three other strategies:

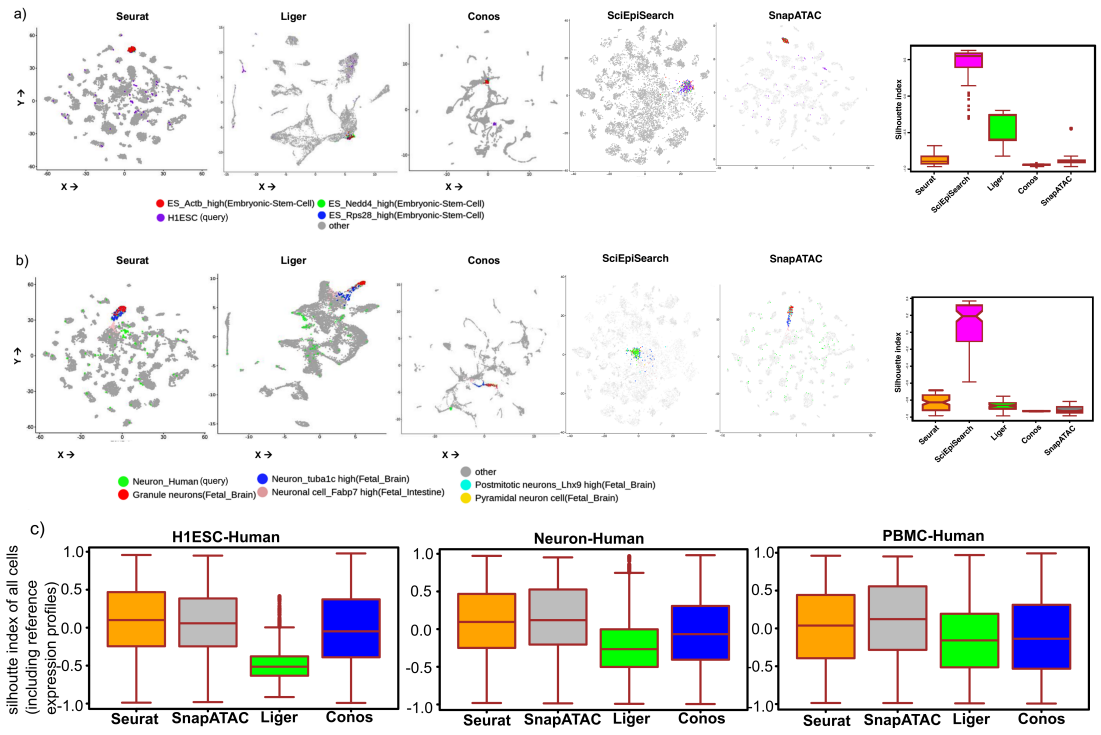

Figure 3.5: Utilizing the reference scRNAseq of mouse cells as the basis for comparison and the scATACseq profile of human cells as the query A) The search comprised of human embryonic stem cell single-cell ATAC-seq profile (H1ESC). Additionally provided are the query cell silhouette coefficients for the four approaches. B) The scATACseq of human neuronal cells were query. Also included are the silhouette coefficients for the query cells. C) The silhouette coefficients presented for all cells in 2D embedding plot, including the reference MCA cells. According to the labels on the corresponding sub-figures, H1ESC, human neurons, or human PBMC were the source of the single-cell ATAC-seq profiles. (Mishra et. al. (248))

- Using scATACseq gene activity scores to compare to gene expression
- scATACseq gene enrichment scores and the reference expression profile correlation based comparison
- Computing the correlation between the reference expression and the predicted expression of the query scATACseq profile based on BABEL (Figure 3.2A).

We discovered that the direct comparison (or correlation) of reference gene-expression values to gene-activity, gene-enrichment, or projected expression of query scATACseq profiles is significantly inferior to our MExTEG-based technique (Figure 3.2A).

Comparison with Integrative methods

We also compared Seurat, LIGER, and Conos integrative approaches, to scEpiSearch in order to identify the single-cell expression profile that most closely matches a spe-

cific type of scATACseq read-count matrix. The identical reference single-cell expression pool of around 1000 cells from the MCA dataset was utilised here (176). We gave Seruat, LIGER, Conos, and scEpiSearch the identical reference cells. We utilised the average of the coordinates (in the tSNE plot) of the top 3 matched cells for the 2D depiction of the scEpiSearch output. We observed that the co-embedding outcomes using integrative approaches with homogenous and small query sets of single-cell open-chromatin profiles were not sufficient (Figure 3.5A,B). It could be because integrative approaches like Seurat and LIGER were not designed to group homogeneous query single-cell ATAC-seq profiles with other non-similar cells incorrectly since they utilise heterogeneity in single-cell profiles to discover anchors. In fact, Leucken et al. (171) pointed out that many integrative approaches' usage of gene activity scores may not be the best way to describe scATACseq data. They also demonstrated the weak performance of integrative methods for single-cell epigenome profiles. Unlike other methods, we did not consider silhouette coefficients for reference single-cell expression data-points in the co-embedding graphs because it might have carried away the corresponding values for query single-cell ATAC-seq profiles, which is another reason for the revelation of such a result (Figure 3.4C). We only generated silhouette coefficients for query cells since our goal was to assess the method of finding matches from expression profiles for single-cell ATAC-seq data sets (like a search engine).

When we included the single-cell ATAC-seq profiles of three different kinds of mouse cells (macrophages, B cells, and endothelial cells), we enhanced the heterogeneity of the query, which resulted in an improvement in the co-embedding plot (Figure 3.3A). However, the integrative techniques (Seurat, LIGER, Conos) were not significant compared to scEpiSearch based on the silhouette coefficient scores for query single-cell ATAC-seq profiles (Figure 3.3B).When human scATACseq profiles were utilised as a query for the same mouse reference expression profiles made up of 10,000 cells from MCA, a similar trend was observed. Human embryonic stem cells (HESC) scATACseq profiles were used as a query, and scEpiSearch based plots revealed they were more similar to mouse ESC. Results based on scEpiSearch for the scATACseq profile of Human Neuron cells revealed their connectivity to reference Neuron Cells from MCA (Figure 3.5B). Whereas Seurat and LIGER had the same issue, causing homogeneous query cells to spread out and colocalize with several clusters of reference expression profiles that were dissimilar in co-embedding plots (Figure 3.5). The outcomes for LIGER

and Seurat enhanced such that query cells appeared proximal to immune cells in their co-embedding plots when we employed a query comprised of human PBMC scATAC-seq profiles with more heterogeneity (Figure 3.5). However, based on the outcomes of scEpiSearch, when we computed the silhouette coefficients for query cells to assess the effectiveness of the integrative technique as a search engine, their performance was not comparable to simple embedding. Overall, our findings show that integrative approaches were unable to effectively search for matching expression profiles for query scATACseq datasets such as scEpiSearch.

Robustness in Accuracy of scEpiSearch : The scATACseq reference expression and gene-enrichment scores of the standalone and webserver versions are stored in their own databases. Scalable visualisation is a feature of both versions of scEpiSearch that enables interactive viewing of more than 1,000 query scATACseq profiles. The majority of the queries were evaluated using multiple scATACseq profiles with known cell-type annotation, and it was found that the search accuracy of scEpiSearch is around 80–100% in identifying the proper cell type among the top 5 hits (Figure 3.2B). We prevent artefacts caused by method or species-specific batch effects by employing proximal gene enrichment and expression profiles rather than relying on correlation or proximity among cells. Consequently, scEpiSearch also enables very accurate comparison of query scATACseq profiles from human cells to mouse reference single-cell expression profiles (Figure 3.2B).

3.3.2 Reference Dependent Graph Embedding of Query sc-ATACseq Profiles

Despite the fact that scEpiSearch can match open-chromatin and transcriptome profiles for a single query cell epigenome, it is often essential to visualise and cluster cellular profiles from several sources in order to get insight into distinct or mixed cellular states. Therefore, regardless of the species of origin, scEpiSearch is also intended to incorporate and give an integrated view of numerous scATACseq profiles with variable peak-list and batch-effects. Based on how closely top-matching mouse reference expression profiles resemble each other, scEpiSearch determines the distances between query cells. When two mouse expression profiles are grouped together in the reference scRNA-seq dataset of scEpiSearch, they are said to be comparable in this context.

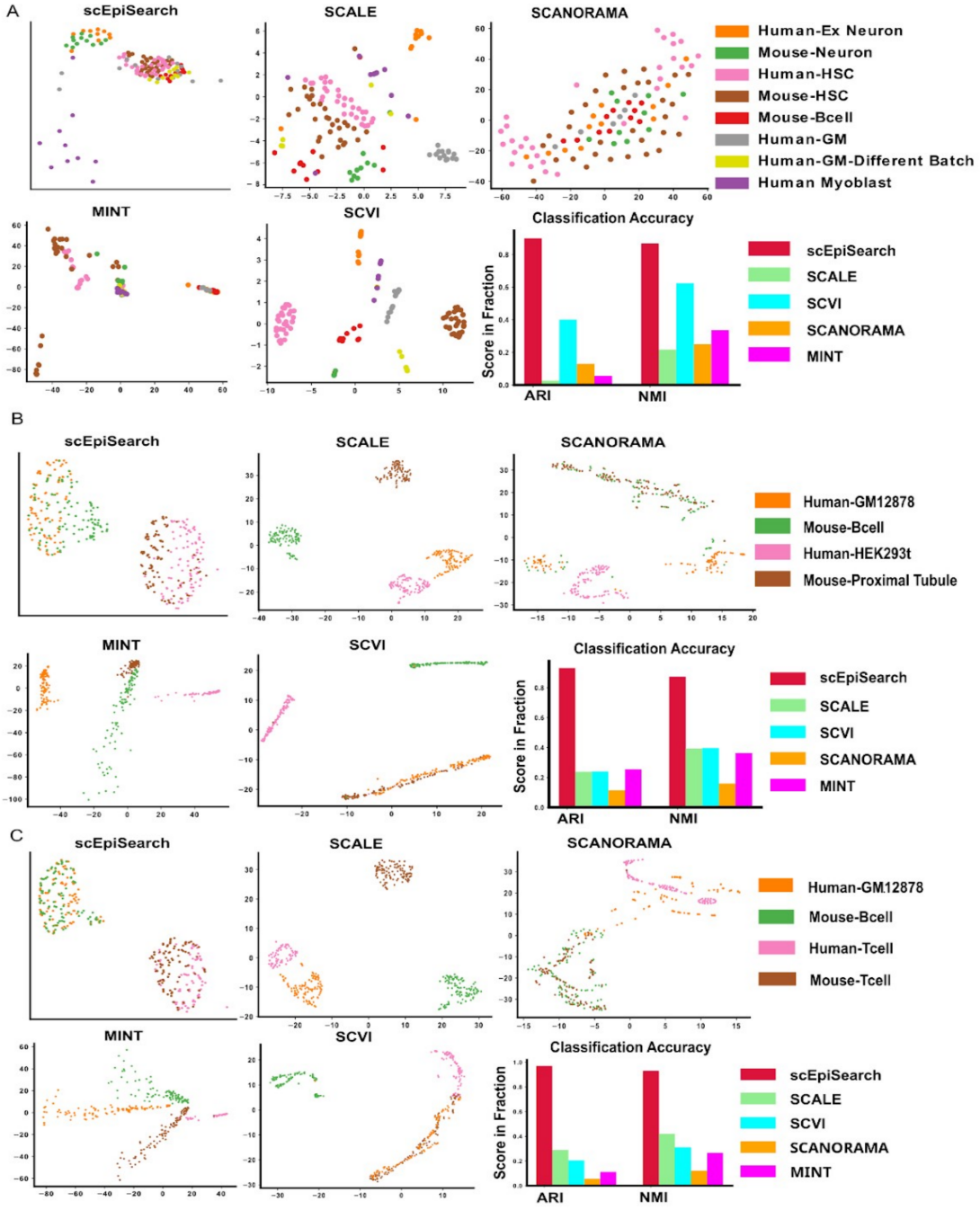

Figure 3.6: Evaluation of embedding of query sets of single-cell open-chromatin profiles irrespective of batch effect, species, differences in peak-list and their source. Embedding plot from ScEpiSearch derived from projections onto mouse expression profiles. (A) Queries consisted of scATACseq profiles of human-neuron, mouse-neuron, human-HSC, house-HSC, human-Myoblast, human-GM12878 cells from two batches and mouse-B-cells. The purity of density based spatial clustering (using DBSCAN) with embedded coordinates is shown here in terms of ARI (Adjusted Rand Index) and NMI (normalized mutual information) scores. (B) Queries made for Human-GM12878 cell, Mouse B-cell, Human-HEK293T, Mouse-Proximal tubule. (C) Queries were made for Human-GM12878, Mouse B-cell, Human T-cell, Mouse T-cell. (Mishra et. al. (248))

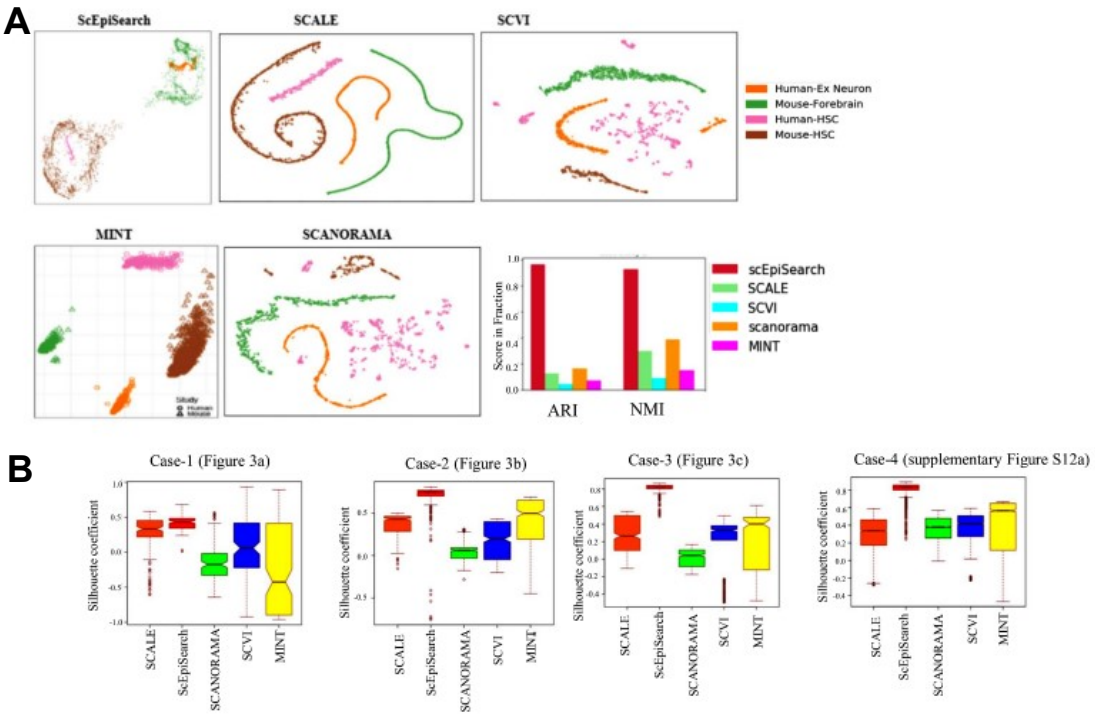

Figure 3.7: Evaluation of query scATACseq profiles from various species and batches using 2D embedding. a) Query for Human-Neuron, Mouse-Neuron, Human-HSC, and Mouse-HSC are made. None of the other techniques examined in this study were able to provide accurate low-dimensional embedding like scEpiSearch. For others,TSNE plot is shown. After utilising labels HSC and forebrain/neurons and applying DSCAN to the 2D coordinates, the right-bottom panel displays clustering-purity in terms of ARI and NMI. b) This plot displays the evaluation of 2D embedding plots showing the computation of silhouette coefficients. (Mishra et. al. (248))

We used four distinct sets of scATACseq read-count matrices to test the efficiency of scEpiSearch with four embedding techniques (SCANORAMA, MINT, SCVI, SCALE) (187) (188) (189) (190). Gene-enrichment scores from various query scATACseq profiles were supplied to SCANORMA, MINT, and SCVI for 2D embedding since they employ genes as features. While for SCALE, 2D embedding was carried out using tSNE and its latent feature space encoding of the scATACseq read-count matrices. The 2D embedding map created by scEpiSearch for scATACseq profiles, as demonstrated in Figure 3.6A,B,C, 3.7A, displays almost accurate colocalization of comparable cell types regardless of the species and research lab of origin. Other embedding techniques (SCANORAMA, MINT, SCVI, and SCALE) produced the incorrect cell grouping (Figure 3.6, 3.7). After density-based spatial clustering (using DBSCAN (191)) of the embedding outputs using cell type labels as true clusters, we evaluated clustering purity to provide further evidence. The ARI (adjusted Rand Index) and NMI (normalised

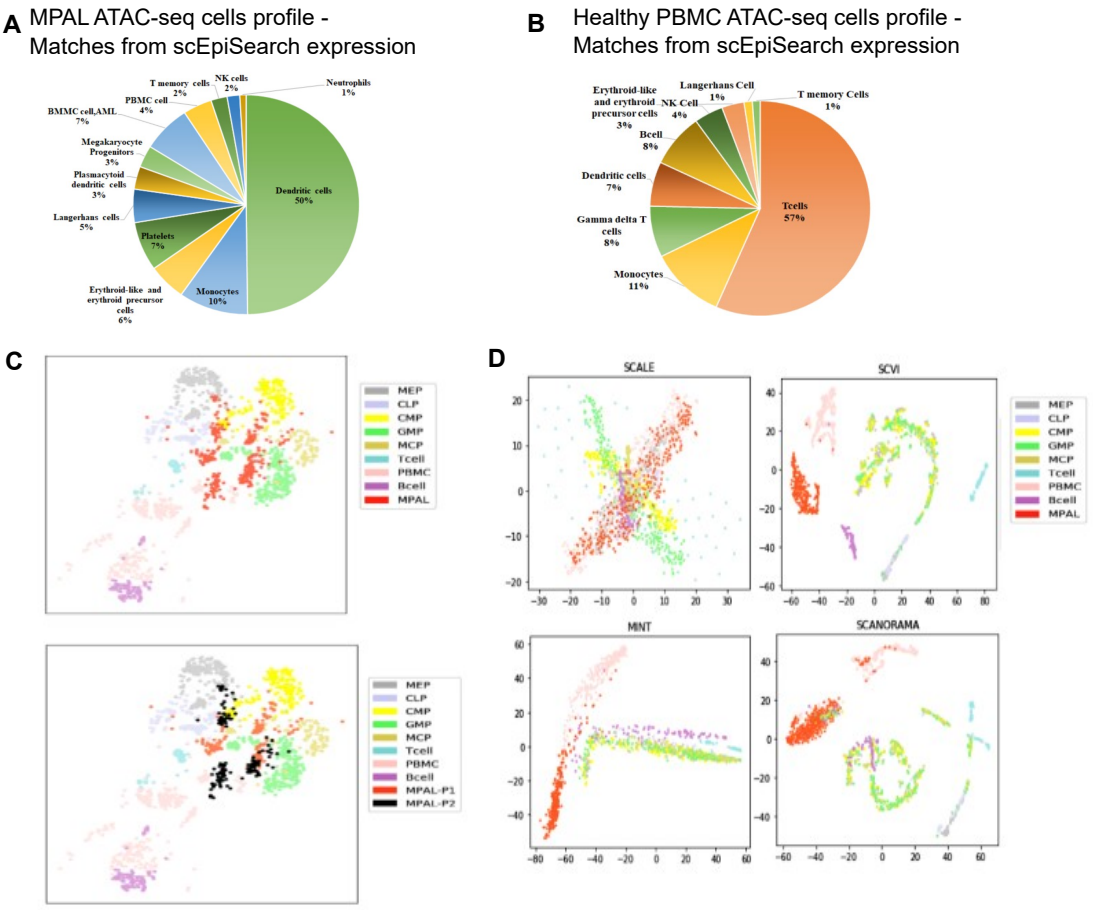

Figure 3.8: Using scEpiSearch's 2D embedding to follow the dedifferentiated state of blood-cancer patients' leukaemia cells. A) Pie chart of top scRNAseq types matches for scATACseq profile of blood cells from individuals with multiple phenotype acute leukaemia (MPAL) is shown in Figure 1. (GEO id: GSE139369). B) Pie-chart displaying cell types of human peripheral blood mononuclear cells (PBMCs) with top matching scRNAseq cells (GEO id: GSE139369) C) scEpiSearch based 2D embedding of scATACseq of 3 kinds of cells: peripheral blood mononuclear cells (PBMC) from healthy cells, progenitors of blood cells, and blood cells taken from patients with mixed phenotype acute leukaemia (MPAL). Granja et al. released the scATACseq profiles of MPAL and PBMC cells, and progenitor cell scATACseq are from a different study. Most PBMC cells are located further from progenitor cells and closer to B cells in the embedding plot created by scEpiSearch. Progenitor cells are further distant from MPAL cells. Some MPAL cells are so thoroughly dedifferentiated that they even entirely overlap with blood cell progenitors. the identical 2D embedding result, but the MPAL cells were coloured in accordance with the origins of the patient (P1 and P2). MPAL cells from two patients have some overlap, but they also have various degrees of dedifferentiation. D) Outcomes from other tools for 2D embedding of scATACseq of three different kinds of cells: peripheral blood mononuclear cells (PBMC) from healthy persons, progenitors of blood cells, and blood cells from patients with mixed phenotype acute leukaemia (MPAL). Other approaches either failed to colocalize B and T cells with PBMC or jumbled up the location of several hematopoietic progenitor cell types. (Mishra et. al. (248))

mutual information) ratings based on clustering purity demonstrated the advantage of scEpiSearch in the objective embedding of open-chromatin profiles (Figure 3.6, 3.7). The comparison using silhouette coefficients is shown in Figure 3.7B.

Case Studies for Embedding of Multiple sc-ATACseq Profiles Across Batches and Species

With the standalone variant of scEpiSearch, we experimented with a number of scATACseq read-count matrix combinations. We were able to successfully incorporate scATACseq profiles, resulting in colocalized cells of the same kind while preserving distinct cell-type separability.

- Case 1 : Cells from Human Neuron (GSE97942), Mouse forebrain (GSE100033), Human HSC (GSE96769), Mouse HSC (GSE111586), Human Myoblast (GSE 109828), Human GM (GSE109828), Mouse B-cell (GSE111586), and Human GM12878 (GSE68103) were taken which had their own batch-effect and belong to both human and mouse species from which we made our queries. We then passed these queries to scEpiSearch to obtain top 10 matching clusters. This information is then used to create an adjacency matrix, which was then forwarded to create the final embedding plot. In order to create graphs and visualize clusters, NetworkX was employed. HSC cells from both species can be seen to be grouped together, whilst neuronal cells out of both species (Mouse and Human) can be seen to be grouped together. The graph shows the separation among HSC and neuron cells (Figure 3.6A). Due to their classification as immune cells, B Cells and GM out of both species may be seen in discrete clusters next to HSC cells. No other type of cells cluster with human myoblast cells. The read-count matrices' peak lists varied from one another. Here are five distinct approaches' embedding charts. While MINT was unable to group cells of the same kind together like scEpiSearch, SCANORAMA was able to mix the locations of all cells.

- Case 2 : Here, we employed read-count matrices from the scATACseq profiles of four different cell types, including the human B cell line (GM12878), mouse B cells, human T cells, and mouse T cells (Figure 3.6B). Human B cells and mouse B cells colocalize in the 2D embedding figure created by scEpiSearch (Figure 3.6B). In scEpiSearch-based embedding, human and mouse T cells colocalize similarly. In comparison, the other four methods were unable to provide such accurate embedding.

- Case 3: We employed four read-count matrices of scATACseq profiles of four different cell types for the third example (Figure 3.6C), including human B cell line (GM12878), mouse B cells, human embryonic HEK293T kidney cell line, and cells from the mouse proximal kidney tubules. In this instance, scEpiSearch offered almost accurate embedding demonstrating the colocalization of mouse B cells with human GM12878 cells (Figure 3.6C). Other techniques, however, displayed incorrect co-localizations in the embedding outcomes, as shown in the

data from the SCALE, where mouse proximal tubule and B cell colocalized. Additionally, SCVI mislocalized mouse proximal tubule and human GM12878 cells.

- Case 4: Human Neuron (GSE97942), Mouse Forebrain (GSE100033), Human HSC (GSE96769), and all mouse HSC from Bone marrow tissue (GSE111586) were selected as samples from different batches and species of larger numbers of cells to demonstrate the consistency of the module. These samples were then passed to the Embedding module of the standalone system (Figure 3.7A). The graph was constructed by creating an adjacency matrix using the top twenty matching clusters, and both groups are clearly separable.

3.3.3 Case Study of Reference based Graph-embedding: Understanding Multiple Phenotype Acute Leukaemia

We examined the scATACseq profiles of patient blood cells with mixed-phenotype acute leukaemia (MPAL) in order to get further insights from the network based embedding of single-cell epigenome profiles and their underlying cell states (192). A change in the proportion of cell types was discovered in the same study's first investigation of MPAL and PBMC cells from healthy individuals. Similar proportions of the various cell types were present in the corresponding expression profiles for the scATACseq profile of PBMC from healthy patients, as reported by others (193) (Figure 3.8A,B). However, we observed an increase in the proportion of dendritic, monocyte, and erythrocyte cell types in MPAL cells from two patient samples (Figure 3.8B). It is not simple to determine if it was caused by sampling bias during scATACseq sequencing or whether it reflected authentic fractions. However, we carried out the embedding of scATACseq profiles of MPAL cells from two patient samples, PBMC from healthy persons, progenitors of hematopoietic cells (progenitors of hematopoietic cells in the blood) (194), T cells, and B cells (195).

The blood's progenitors of hematopoietic cells, including the CLP (common lymphoid progenitor), MEP (megakaryocytic-erythroid progenitor) CMP (common myeloid progenitor), MCP (mast cell progenitor), and GMP (granulocyte-monocyte progenitor) were included in our investigation. Many MPAL cells overlapped with different sorts of hematopoietic progenitor cells in the 2D embedding graph from scEpiSearch. T cells, B cells, and several MPAL cells are grouped with PBMCs (Figure 3.8C). In the findings of our embedding, hematopoietic progenitor cells and PBMC hardly overlapped. These findings provide some indication of the various degrees and variations

of de-differentiated states of MPAL profiles. The dedifferentiated states of MPAL profiles may account for their flexibility and capacity for lineage flipping (196). Only a small number of the MPAL cells for two individuals exhibited overlap, according to more thorough research (Figure 3.8B). The majority of cells from two MPAL patients did not overlap and shown a strong relationship with several kinds of hematopoietic progenitor cells. Additionally, we used SCANORAMA, MINT, SCVI, and SCALE on the similar set of read-count matrices, and we discovered that none of these approaches exhibited PBMC colocalization with B cells or T cells nor did they mix the locations of various blood cell progenitors (Figure 3.8D). Given that B cells and T cells are often found in PBMC, it soon became evident that utilising alternative techniques would not provide the same results as scEpiSearch (Figure 3.8C,D). In conclusion, scEpiSearch can highlight dedifferentiated states and adaptability of cells originating from patient samples and may show similarity and differences across subpopulations of cells regardless of source and batch impact.

3.4 Discussion

Projection of single-cell epigenome profiles accurately and robustly onto large pool of reference expression profiles and interpret matching cells in a meaningful way is a challenge. Technical and batch biases, such as cell-to-cell variability in signal i.e peak accessibility and noise, differing read-depths, methods, platforms, and labs, make this problem difficult to solve. The cross species matching of single-cell epigenome profiles was made possible by the computational approach we proposed in this work. Notably, in order to minimize batch and technical biases, our technique relies on median expression and enrichment of top genes and peaks (MEXTEG), rather than distance-based methods such as correlation or cosine distance, hashing, or latent-feature extraction methods. Because of scEpiSearch's large pool of reference cells and statistical method for reducing bias, scEpiSearch is distinctive in its ability to find equivalent transcriptome and open-chromatin profiles. In addition to facilitating single-cell searches against a large collection of reference cell profiles, scEpiSearch also offers low-dimensional graph based embedding. Therefore, the cross-validation of rare cellular states identified by single-cell open chromatin profiles may also be accomplished using scEpiSearch. The standalone version may be securely, and locally for sensitive and clinical data and

includes an inbuilt reference set for processing.

Other Applications of Reference based Matching for improving graph based embedding are: Reference-based matching has a variety of potential implications, including

- Identifying the appropriate cell types from in vivo samples of unannotated cells using scATACseq profiles.
- Analyzing heterogeneity and monitoring changes in the cell state and their potency or divergence, such as as one cell line exhibiting the ability to differentiate into many lineages, such as K562, HL60 cells (177).
- Identifying a suitable mouse model for an imprecise human cell.
- Numerous scATACseq profiles, regardless of their sources, protocols, or species, can be embedded and clustered. To illustrate such applications, we have used ScEpiSearch for four different kinds of case studies below.

As shown in the network based co-embedding of scATACseq and single-cell expression profiles, our research demonstrated the advantages of mapping single-cell epigenomes to reference-cell profiles. Results based on knowledge oriented reference based edge weight learning surpassed integrative approaches like Seurat, LIGER, and Conos. The method using reference based graph embedding approach also revealed that embedding of open-chromatin profiles with distance calculated using projection on reference cell profiles could be better than distance estimated by SCALE, MINT, and SCANORAMA or other latent reference-free feature extraction techniques. Additionally, by using the reference-free feature extraction, novel features in minority populations of cells could be masked by characteristics of cell-states in the majority.

The developed method introducing a Graph based approach helps learn underlying distance among cells from various species/batches with different peak lists. The knowledge based learning of edge weights between cells imparts tremendous improvement in joint embedding of cells that includes a novel approach for computing the distance between cells by projecting them onto reference pool of dataset. This assists with the elimination of batch effect and obtain a reliable cell neighbourhood.

This approach of utilizing knowledge from reference profiles to jointly embed query datasets yields meaningful patterns in diseases profiles. The method also demonstrates applications such as knowledge based edge-weights learning provides findings indicating towards various degrees and variations of de-differentiated states of MPAL profiles.

The de-differentiated states of MPAL profiles may account for their flexibility and capacity for lineage flipping.

Furthermore, reference based co-embedding demonstration helps shows distinct phase of cancer cell's de-differentiation and anticipate a cell's behaviour in an unknown condition. The approach makes use of the additional information found in the reference pool of cells in various cellular-states, such an approach has the ability to result in improved annotation and regulatory-inference from query single-cell open-chromatin profiles.

Thus, reference based edge-weight learning yields co-embedding and demonstrates other applications such as understanding de-differentiated states of cancer cells, emphasising imprints within individual cells that are a sign of apoptosis and a stress response in a subpopulation of embryonic stem cells, evaluation of enhancer landscape activity in single cells, the incorporation of more single-cell histone modification datasets and provision of a powerful reference based search for query cells for key epigenetic regulators.

CHAPTER 4

Explainable Predictive Model using Graph-Wavelet for Modelling Biophysical Properties of Proteins and Measuring Mutational Effects on Diseases

With advancement in experimental and computational methodologies for protein structures, there is breakthrough in understanding more about the connections between protein structure, sequence, dynamics, and function attributable to the tremendous proliferation of data on protein structure and sequence. This information is used to assess the relationships between proteins and other compounds, peptides, and possible therapeutic targets (198). Numerous experimental and theoretical techniques have been used to study protein structures and their functions (199) (200) (201) (202). Recently, researchers have looked at the computational method known as network analysis, which transforms the protein structure into a network/Graph of amino-acid residues. In the field of protein systems research, a novel paradigm has been established by the network-based exploration of interactions (203) between amino acids in protein structure. Protein architectures are represented by residue interaction networks, which are undirected networks of amino acids and their relationships (204) (205). Residue interaction networks disseminate topological information and have the ability to expose the general characteristics of a protein molecule, according to Vendruscolo et al. (206) and del Sol & O'Meara (207).

According to Dokholyan et al. (208), certain network characteristics are associated with protein unfolding, folding rates, and several key residues that serve as the nucleation for protein folding exhibit significant betweenness values in protein transition states. The effect of edge removal per residue, calculated as the percentage of change in average path length divided by protein size to edge removal probability, has been shown by Jung et al. (209) to have a substantial association with protein unfolding rate. The network parameter coefficient of assortativity has a positive association with protein-folding rates, according to Bagler & Sinha's study (210). In order to discover

crucial residues for protein function, Cusack et al. (211) calculated the frequently visited residues in networks based on shortest distance and betweenness. Recently, protein characteristics have also been modelled using graph convolutional networks (212).

Genetic involvement is a crucial driving force in organism evolution. Particular genetic mutations, like SNPs, can be harmful and result in disease in people. Single amino acid polymorphisms (SAPs) are SNPs that cause mutations in amino acids and are thought to be primarily linked to a variety of disorders. However, so far, only a small percentage of SAPs have been connected to ailments. Additionally, it has been demonstrated that using network topological properties, disease-related single amino acid polymorphisms can be predicted to some extent. According to Li et al. (213), because such aberrations often occur at residues with significant centrality or degree, nearby residues of a mutation site may aid in determining if a mutation is disease-related. Using network parameters, protein's functional modules have indeed been researched. In order to identify functional residues and functional module clusters in rhodopsin that encode classic coevolutionary information in the protein's amino acid network, Park Kim (214) used structure-based correlation mutation analysis.

A mechanistic understanding of the deleterious effects of amino acid mutations may be gained by coupling their property to the behaviour of proteins. So, using machine learning methods, we created a novel method for analysing the biophysical characteristics of proteins that is based on graph wavelets. The main benefit of the graph wavelet transform would be that it fractionates the signal into a pattern of scores with multispectral resolutions that match to the specific details of latent modules in the graph. Our approach integrates multispectral data of network-based and physicochemical characteristics of amino acids (nodes) and their interactions in the protein structure. In this article, we provide a model for estimating the rate of protein folding and distinguishing transmembrane-globular, soluble/insoluble proteins, and -helices/-strands. Our data was collected from RCSB/PDB (217), which covers the majority of protein structures. Additionally, it aids in determining how mutations affect multispectral feature rankings, which may be utilised to account for its negative impact. Therefore, we go on to demonstrate how the trained machine learning model gains explainability by estimating the significance of multispectral characteristics.

Hence in our third contribution, we introduced a novel methodology that leverages

graph signal processing to predict biophysical properties of proteins. This approach utilizes graph wavelets derived from physicochemical signals associated with amino acids in protein residue networks to model and predict the biophysical properties of proteins. By utilizing graph wavelets, we were able to estimate the potential effects of disease-associated mutations on these properties. Through our predictions, we provided valuable insights into how specific mutations may impact the overall structure and folding dynamics of proteins. The application of graph signal processing and graph wavelets in protein biophysics has significant implications. It enables researchers to gain a deeper understanding of the structure-function relationship in proteins and offers a valuable tool for predicting the effects of mutations on biophysical properties.

4.1 Material and Methods

4.1.1 Weighted RIG Model

Our approach generates a graph $G(V, E)$ of each protein molecule separately, with each amino acid represented by a node in the graph and the distances between them represented by an edge. Typically, a residue interaction graph(RIG) is used to simulate a protein's network of residues. An adjacency matrix is created in a RIG-based network according to the criteria i.e two residues (vertex) are linked if their physical distance is less than or equal to a predetermined threshold. We have used a weighted RIG-based network that is capable of incorporating both long- and short-range interactions. Since they are responsible for maintaining the protein's structural integrity, long-range interactions are crucial for predicting protein function. The three-dimensional coordinates of the atoms in the protein structure found in the Protein data bank (PDB) files are used to compute the distance between residues. Our definition of the residue's centre is its alpha carbon atom, or $(C\alpha)$. The physical distance $(d_{i,j})$ between the residues (vertices) Vi and V_j is computed using the $C\alpha$ coordinates (x_i, y_i, z_i) and (x_j, y_j, z_j) in the PDB structure.

$$d_{i,j} = (x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2 \tag{4.1}$$

If this computed distance is less than or equal to a predetermined threshold, (rc), then these vertices are regarded as being connected. In this case, we used $rc = 8$. The weighted RIG model gives an edge between two residues a value that is inversely proportional to the distance in the sequence between the two amino-acid (or residue) sites. The weighted RIG model's edge weights are computed as

$$weighted\, RIG(i,j) = |j-i|\; if\; d_{i,j} <= \; r_c, otherwise\, 0 \tag{4.2}$$

We discovered via our internal research that the weighted RIG model often performs better than the boolean RIG model.

4.1.2 Amino Acid Features: Curation and Extraction

For each residue in the network of residues in the protein that was modelled using a residue interaction graph, node degree was computed (RIG). Similar to this, node weighted degree was calculated using the network of residues while also taking the weight of the edges and the degree of the nodes into consideration (residues). The extent to which nodes in a network tend to cluster together is quantified by the clustering coefficient. It was calculated for the network's nodes (also known as residues) using Python's clustering() method from the networkx package. The network created using a weighted RIG matrix has a frequency/count of residues that correlates to it. The AAindex database (221) was used to assemble scores for each amino acid's hydrophobicity, bulkiness, turn tendency, molecular weight, flexibility, partial specific volume, coil tendency, compressibility, and refractive index, polarity which were then used to compute feature scores for each protein sequence. Using the source code from Capra JA et al. (245), which enables file format conversion and employs Jensen-Shannon divergence to calculate conservation scores for proteins, the conservation score was calculated by converting pdb files to fasta.

- Hydrophobicity: The hydrophilic and hydrophobic properties of each of the 20 amino acids have been taken into consideration in the development of a hydropathy scale. The scale is associated with a number of experimental discoveries that were read about in the literature (241).
- Turn tendency: A turn tendency is a structural motif when the distance between two residues' $C\alpha$ atoms is small, often between one and five peptide bonds and less than seven (0.70 nm) (244).

- Coil tendency: In proteins, specific hydrophobic, positively and negatively charged residues have been shown to promote the formation of additional coiled coil domains (243). On the other hand, the surrounding hydrophobicity of residues in coiled-coil domains is much lower than that of residues in other coiled-coil protein regions. As a result, it becomes a crucial factor to take into account when modelling the biophysical characteristics of proteins.

- Flexibility: Protein structural flexibility is essential for allostery, binding, and catalysis. Flexibility was inferred from the amino acid sequence using a sliding window averaging method, and it was largely used for epitope search. Based on whether or not its neighbours are stiff or flexible, every amino acid in the protein chain is given a flexibility parameter (240).

- Bulkiness: Bulkiness, which measures the chain's average cross section, is defined as the side chain's volume to length ratio (236). The bulkiness values were determined using the data of Waugh et al. (237) and Sorm et al. (238).

- Partial specific volume: An experimental measure that provides data on solute-solvent interactions and protein hydration is an amino acid's partial specific volume (239).

- Compressibility: According to the Miceller model compressibilities (235), the equation for a protein's compressibility, K_m

$$K_m = \alpha N(CH_2)\, 8.0\, X\, 10^{-15}/M \tag{4.3}$$

  Where M is the relative molar mass, $N(CH_2)$ is the total number of equivalent methylenes, and α is a parameter used to determine the folding pattern.

$$A = (3/4\pi N)[(n^2 - 1)/(n^2 + 2)] \tag{4.4}$$

  Here the number of molecules per unit volume is denoted by N.

4.1.3 Spectral Graph Theory

A weighted undirected graph G composed of an established set of vertices V and a set of positive weights $w : E \implies R$ assigned to the graph's edges E. Let $N = |V|$ be the equation for the number of vertices. For spectral graph theory, the weighted adjacency matrix (NXN) produced from the weighted RIG model serves as a finite weighted graph. The sum of the weights of the edges impacting on the vertex is used to determine the degree of vertices in a graph. It is also possible to define the degree of the vertex m as $d(m) = \sum_n A_{m,n}$ in respect with adjacency matrix. A diagonal degree matrix would thus be:

$$D(i,j) = d(i)\, if\, i = j,\, otherwise\, 0 \tag{4.5}$$

The equation $L = D - A$ describes a non-normalized Laplacian operator for the graph. Instead, we have employed the Laplacian operator's normalised form as provided by:

$$L^{norm} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2} \tag{4.6}$$

The diagonalization of the matrix is possible due to its symmetry, which leads to the eigenvalue decomposition $L = U\lambda U$, where U is the eigenbasis matrix and λ is the diagonal matrix containing L's eigenvalues. The orthonormal eigenvectors of the eigenbasis matrix U are designated as χ_l for $0 \leq l \leq N - 1$. The corresponding eigenvalues λ_l fulfil the equation $L_{\chi_l} = \lambda_l \chi_l$. Given that G is linked and given that L is symmetric, it is likely that the eigenvalues are real, non-negative numbers that may be ordered in ascending order as $0 = \lambda_0 < \lambda_1 \leq \lambda_2 ... \leq \lambda_{N-1}$.

4.1.4 Spectral Graph Wavelet (SGWT)

According to SGWT's (218) definition, it acts as a bandpass filter and fixes a non-negative real-valued kernel function called g. Additionally, $g(0) = 0$ and $\lim_{\lambda \to +\infty} g(\lambda) = 0$ are necessary. The wavelet operators obtained as the rescaled kernel function of the graph Laplacian provide the SGWT coefficients at each scale. The eigenvectors and eigenvalues of the Laplacian matrix (L), on the other hand, may do this for a graph with a Laplacian in a finite dimension. The wavelet operator is specifically specified by $T_g = g(L)$. The function $T_g f$ returns wavelet coefficients at scale $(s) = 1$ for a signal f. Depending on what this operator does with the eigenvectors χ_l, it is described as

$$T_g \chi_l = g(\lambda_l)\chi_l \tag{4.7}$$

As a result, the operator may be stated to operate on the graph signal f by adjusting each graph Fourier coefficient as

$$Tgf(\hat{l}) = g(\lambda l)\hat{f}(l) \tag{4.8}$$

The wavelet operator then would be $T_g^s = g(sL)$ for each scale. This is described in terms of the continuous $(g(\lambda))$ domain of the kernel function. Furthermore, $\psi_{s,n} = T_g^s \delta_n$ would be the spectral graph-wavelet at scale s, centred on vertex n. Since f and $\psi_{s,n}$ are inner products of each other, wavelet coefficients may be thought of as such

$$W_f(s,n) =< \psi_{s,n},\ f > \tag{4.9}$$

4.1.5 Implementation Details

Building a weighted RIG model of each protein is the first step in our technique (ProteinGW) for predicting the properties of proteins using graph wavelets , which then uses signals on the graph to represent the physiochemical, network-based, and conservation properties of amino acids (Figure 4.1). Molecular weight, amino acid count frequency, conservation score, hydrophobicity, polarity, refractive index, turn tendency, coil tendency, partial specific volume, compressibility, flexibility, bulkiness and some graph properties like node weighted degree, clustering coefficient and node degree are signals for the amino acids that have been acquired. The number of edges that are close to the vertex in this case determines the node degree, and the edge weights of the edges that incidence on the vertex determine the node weighted degree. The geometric average of the edge weights of the subgraphs, or the clustering coefficient

$$C_u = \frac{1}{deg(u)\ (deg(u)-1)} \sum_{vw} (\hat{w}_{uv} \hat{w}_{uw} \hat{w}_{vw})^{1/3} \tag{4.10}$$

where, C_u tends to 0 when $deg(u) < 2$ and $(\hat{w}_{uv})$ (edge weights) are normalized by network's maximum weight i.e., $\hat{w}_{uv} = w_{uv}/max(w)$.

By default, ProteinGW produces wavelet coefficients at four different scales. During the model's evaluation, ProteinGW does hard thresholding by determining the best percentile threshold for those wavelet coefficients. ProteinGW divides the dataset into training and test sets, then uses 5-fold cross-validation to train machine learning models. ProteinGW employs Accuracy, ROC AUC, Macro-F1, and MCC scores as evaluation metrics for classification tasks and R-sqaure, RMSE for assessment of protein-folding rate while evaluating the prediction model. Here, we compare random forest with sev-

eral machine learning models for the classification tasks of transmembrane/globular, all-α/all-β, soluble/insoluble, XgBoost, AdaBoost, K-Nearest Neighbors, Gaussian Naive Bayes, Logistic Regression, and SVM. Additionally, we contrasted the random forest regressor with ElasticNet, Decision Tree, K-Nearest Neighbors, SVR, Ridge, Lasso, and Linear regressions for protein folding rate estimate using the ProteinGW method. Furthermore, ProteinGW determines feature importance at each scale for all physico-chemical/network properties to increase explainability to the trained Random forest prediction model. Finally, it is also investigated how top predictive traits change at protein mutation locations that cause certain diseases. We investigated the relationship between amino acid network characteristics and disease-associated mutations using publically accessible materials (219).

4.1.6 Other Graph Signal Based Methods

Graph Fourier Transform (GFT)

Both the network characteristics and the properties of the residue, which were features employed in ProteinGW, were the signals that were selected. For each of the three characteristics (alpha-beta, solubility, and transmembrane globularity), a linear multivariable classification model was developed, along with a regression model for protein folding rate (231). The frequency cutoff for each signal was chosen to optimise the correlation between the signal's low-frequency component and the protein biophysical feature that was being modelled.

Convolutional Neural Network

Given a graph's adjacency matrix, the matrix may transform it into a 2D picture with many channels, much like a traditional image representation format, which consists of a 2D matrix that denotes the pixel densities and several channels that, in most instances, represent RGB channels. Therefore, we employed eigenvalues and eigenvectors to translate an adjacency matrix of a network into its counterpart in an image format. The adjacency matrix's eigenvalues and related eigenvectors were first taken out. We then removed the 'n' biggest eigenvalues and eigenvectors that corresponded to those values following this operation. We normalised the 'n' eigenvectors after extraction such that

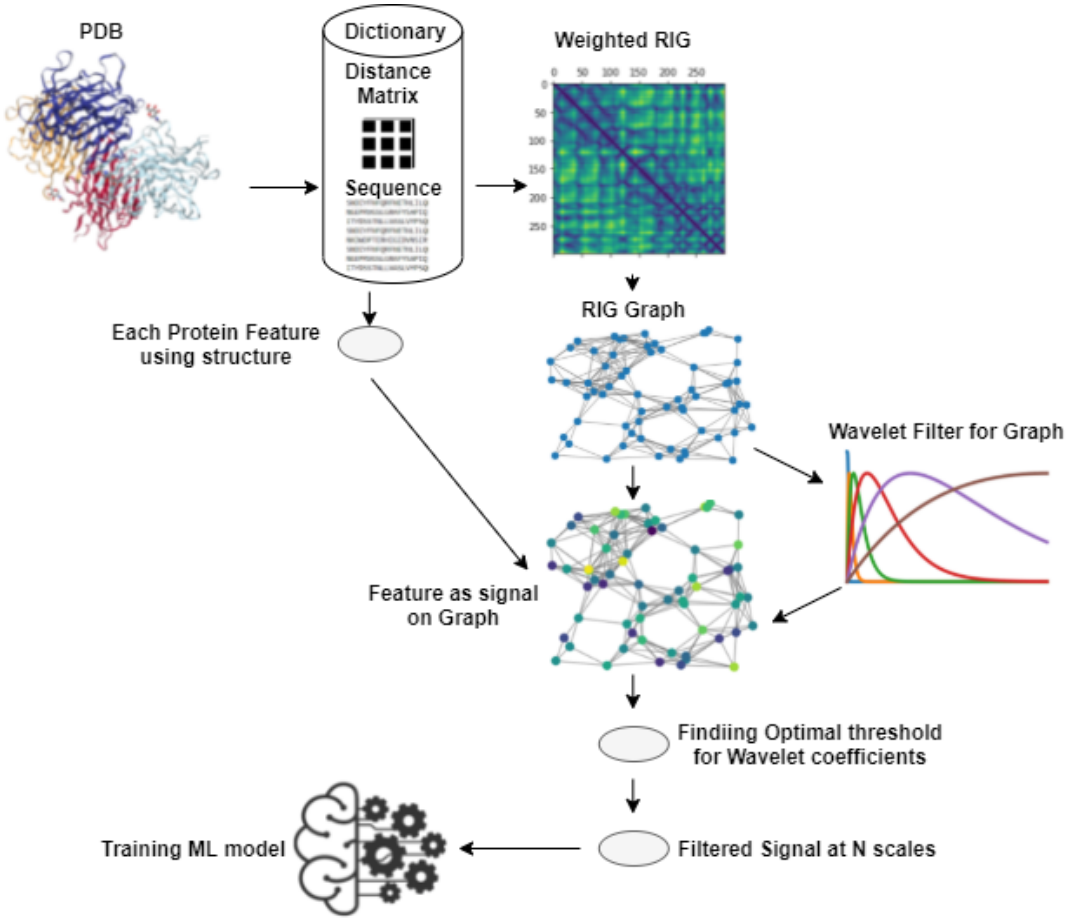

Figure 4.1: The pipeline for ProteinGW is shown in a flowchart. For the modelling of each protein function, protein structures were acquired from PDB database. For each protein, a dictionary was made that included the distance matrix, sequence, etc. Adjacency matrix known as weighted RIG was constructed. Then, using a weighted-RIG matrix, a network tailored to each protein is constructed. Additionally, each amino acid (node) is represented on the graph as a signal with its physical, chemical, and network characteristics, and the ideal threshold is determined to cutoff wavelet coefficients. An ML model is then trained. (Mishra et. al. (249))

their values fall inside a predetermined range between [-1,1]. We initialised a blank three-dimensional matrix with the shape (n/2,28,28), where n/2 denotes the number of channels and 28,28 denotes the shape of a single channel, in order to represent these eigenvectors as images (resolution) (232).

After completing this phase, we chose two eigenvectors at a time, one representing the dimension x and the other dimension y, and filled the values in the corresponding pixels, just as we would do in a histogram bin. Therefore, we effectively split our (28,28) matrix such that a value would go into a given cell of the matrix if it was located in a certain area. By selecting eigenvectors in pairs, values were filled in all n/2 channels in this manner, and conventional CNN was then applied to the resulting picture

representation. For CNN, we started with a kernel of form (3,3), added a max-pool 2D layer, flattened the features, and then added four dense layers using the activation functions "relu", "linear", and "softmax". The first three thick layers had 100, 50, and 20 neurons, respectively, while the number of neurons in the last layer varied depending on the task at hand.

4.1.7 Data Sources

52 single domain two-state folding proteins were employed for the modelling folding rate that we used. Multiple sources were used to compile the data for those proteins (200) (220). The alpha/beta modelling process employed the same 52 proteins. While we gathered 237 transmembrane and 59 globular proteins from the PDB library to mimic the transmembrane/globular feature. We utilised the proteins on the list provided by Han et al. that have a PDB structure to predict solubility.

4.2 Results

Each amino acid has a unique set of physicochemical properties that affect how proteins behave generically and how quickly they fold. Therefore, while comparing proteins and investigating their functions, it is required and prudent to extract characteristics based on amino acid properties. We utilised the hydrophobicity scores, flexibility, polarity, relative molecular mass, isoelectric point,ionisation equilibrium constant (221), and conservation (245) of 20 amino acids, which are known physicochemical properties. Additionally, we estimated the protein folding rate of proteins and classified proteins based on their transmembrane/globular, soluble/non-soluble, alpha/beta, and network characteristics of amino acids in the residue interaction graph (RIG) (Figure 4.1). ProteinGW uses a wavelet transform to position signals for each feature on the weighted RIG network of each protein. Additionally, it produces overall multispectral feature scores that may be employed in machine learning approaches after determining a consensus cutoff to remove noise and irrelevant components from wavelet signals. In all, ProteinGW generates 60 feature scores, with 4 feature scores for each amino acid characteristic, each of which corresponds to a different wavelet resolution level.

4.2.1 The predictive Power of Graph-Wavelet Based Feature Extraction

Transmembrane-Globular: First, we predicted the Transmembrane-Globular characteristic of proteins using our method. Globular proteins perform a variety of three-dimensional tasks, including catalysis, transport, cellular signalling, and others. The orientation of a transmembrane protein in the membrane, however, is usually highly particular. The amino acid residues with nonpolar side chains that split these domains make up the majority of the membrane-spanning regions of the polypeptide chain that interact with the lipid bilayer's hydrophobic environment. All peptide linkages are compelled to form hydrogen bonds since they are polar and the bilayer is devoid of water (223). We gathered 59 globular proteins and 237 transmembrane proteins to classify globular and transmembrane using graph wavelets. Following the procedure described in the techniques section, multispectral feature scores, also known as graph wavelet-based feature scores (GWFS), were computed. Additionally, we analysed the performance of several machine learning models utilising GWFS. Figure 4.2A displays several categorization assessment metrics including accuracy, Macro F1 score, ROC-AUC (Receiver Operator Characteristic - Area Under Curve), and MCC (Mathews correlation coefficient). These results were deduced after applying 5-fold cross-validation. The figure clearly shows that the accuracy of the random forest and XgBoost based classifications are superior to other models, with scores of 0.93, 0.88 Macro-F1, 0.86 ROC-AUC, and 0.77 MCC. Hence, we used their results for further analysis.

Solubility: We further attempted to use our method to simulate proteins' solubility. High levels of protein expression and solubility are necessary for successful recombinant protein synthesis, although these requirements are sometimes challenging to meet. Protein research is hampered by several mistakes made during recombinant protein synthesis, particularly structural, functional, and pharmacological studies that are needed for soluble and concentrated protein samples (224) (225). Therefore, it is a contested issue in research to predict solubility and modify protein sequences for enhanced solubility. Protein solubility is influenced by a variety of intrinsic variables, including molecular weight, amino acid content, and residue hydrophobicity (226) (227). In order to predict the solubility of proteins, we employed GWFS derived using such amino-acid characteristics in combination with machine learning classifiers. Random Forest out-

performed other machine learning algorithms in performance assessment using 5-fold cross-validation, with classification accuracy of 0.79, 0.78 Macro-F1, 0.79 ROC-AUC, and 0.57 MCC scores (see Figure 4.2C).

All-α and All-β: Using our method, we also aimed to tackle the problem of quantifying the content of all-α and all-β in proteins. Globular proteins may be divided into four structural categories: all-α, all-β, $\alpha + \beta$, and α/β. all-α and all-β proteins are categorised as being essentially entirely made up of α-helices and β-strands. For the composition of $\alpha + \beta$ proteins, distinct segments of α-helices and β-strands (primarily antiparallel) were recognized, while mixed segments of α-helices and β-strands were described for α/β proteins (mainly parallel). We investigated how well our wavelet-based learning algorithm performed when classifying proteins that are either alpha or exclusively beta. After determining the ideal threshold, classification of α-helices and β-strands is also accomplished by extracting features in the wavelet domain with frequency cutoff using hard thresholding. We mostly observed the performance of all machine learning models, which were all extremely similar in terms of accuracy, macro-F1, ROC AUC, and MCC scores (features from ProteinGW are utilised) (Figure 4.2B). The highest result, however, displayed values of 0.75 accuracy, 0.73 Macro-F1, 0.75 ROC AUC, and 0.57 MCC.

Protein Folding rate: We utilised ProteinGW to predict the protein fold-rate after testing the effectiveness of our technique on three protein-classification tasks. An indicator of how rapidly (or slowly) a protein folds from its unfolded state to its natural 3D structure is the rate of protein folding. Research on protein folding rates aids in a better comprehension of variations in protein folding kinetics that may have a role in diseases like prion and Alzheimer's disease, among others (228). We gathered protein structures (229) and their protein folding speeds in order to assess our approach for modelling protein folding rate. ProteinGW was used to engineer their characteristics. Regression models were developed using the training dataset, and 5-fold cross-validation was used to assess performance using the test dataset. Figure 4.3 displays the RMSE and R-value of several regression models, including linear, lasso, ridge, SVR, KNN, random forest, decision tree, and elastic net, that were tested using features derived from our technique. Figure 4.3 clearly shows that random forest, with R= 0.81 and RMSE=1.68, has outperformed all other models. We discovered that the performance of the other three approaches (Pred-PFR, FoldRate, and SWFoldrate) was worse

Method	R	RMSE
ProteinGW	0.81	1.68
Pred-PFR	-0.04	3.86
FoldRate	0.17	4.64
SWFoldrate	0.93	5.81
GFT (Graph Fourier)	0.74	1.92

Table 4.1: R and RMSE for modelling Protein folding rate for various methods

to ProteinGW when we evaluated the RMSE for folding rate prediction offered by those three other methods (Table 4.1).

4.2.2 Comparing Graph Wavelet with Other Graph-Signal Based Methods

We evaluated our methodology with others that simulate the biophysical characteristics of proteins using protein structure based networks and properties of amino-acid on nodes. One such technique was recently published (231) and is relied on the graph Fourier transform. The other technique we employed for comparability is based on convolutional neural network, which similarly employs graph structure to aggregate signal from an input vector of measured amino acid property. Deep learning based Convolutional Neural Network (CNN) has achieved great success in the area of protein function modelling, however it requires a large number of data samples to train a network of deep learning. In general, obtaining a large enough sample size for training is too challenging, and given the limitations of a limited dataset, overfitting is most prone to occur.

However, we may utilise the Fourier transform to determine the signal amplitudes needed to accurately recreate any signal (231). The Fourier transform does have the fundamental restriction that all signal properties must be considered globally. However, wavelet transform effectively accesses localised signal information by representing the signal in both the temporal and frequency domains (218). In order to assess the robustness and consistency of these approaches, we examined the performance of CNN, GFT, and ProteinGW when a smaller fraction of the training dataset was used.

In order to assess the performance of the model when trained with fewer data

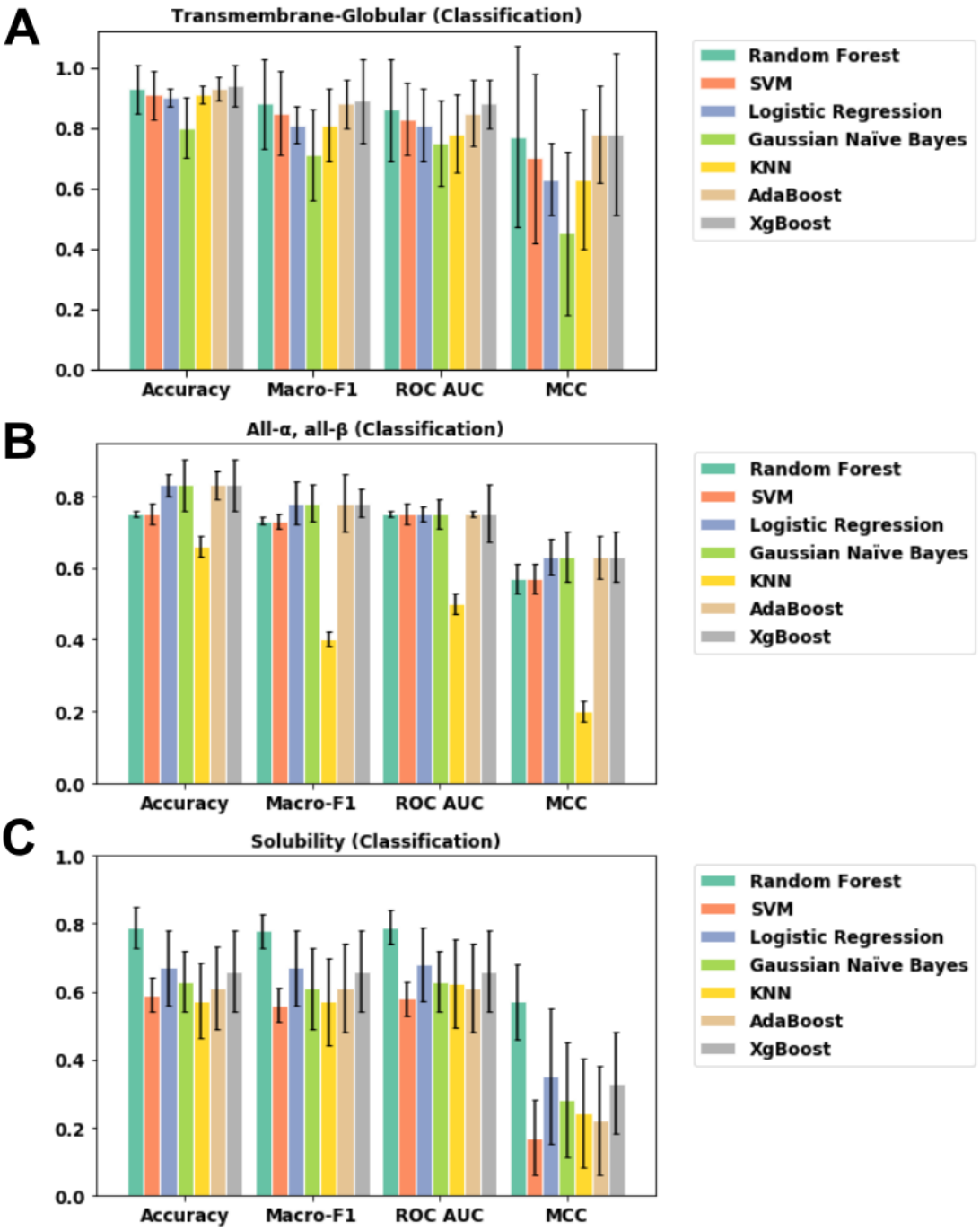

Figure 4.2: Performance of several machine learning models after feature extraction using the graph-wavelet method utilising five fold cross-validation. For the purpose of modelling the categorization of transmembrane and globular proteins, (A) MCC, Accuracy, Macro-F1, and ROC-AUC are compared. XgBoost, AdaBoost, KNN, Gaussian Naive Bayes, logistic regression, SVM, and random forest are machine learning models that are compared.For modelling the classification of soluble and insoluble proteins by machine learning models, (B) MCC, Accuracy, Macro-F1, and ROC-AUC are compared. (C) MCC, Accuracy, Macro-F1, and ROC-AUC are compared for modelling classification of all-alpha and all-beta proteins. (Mishra et. al. (249))

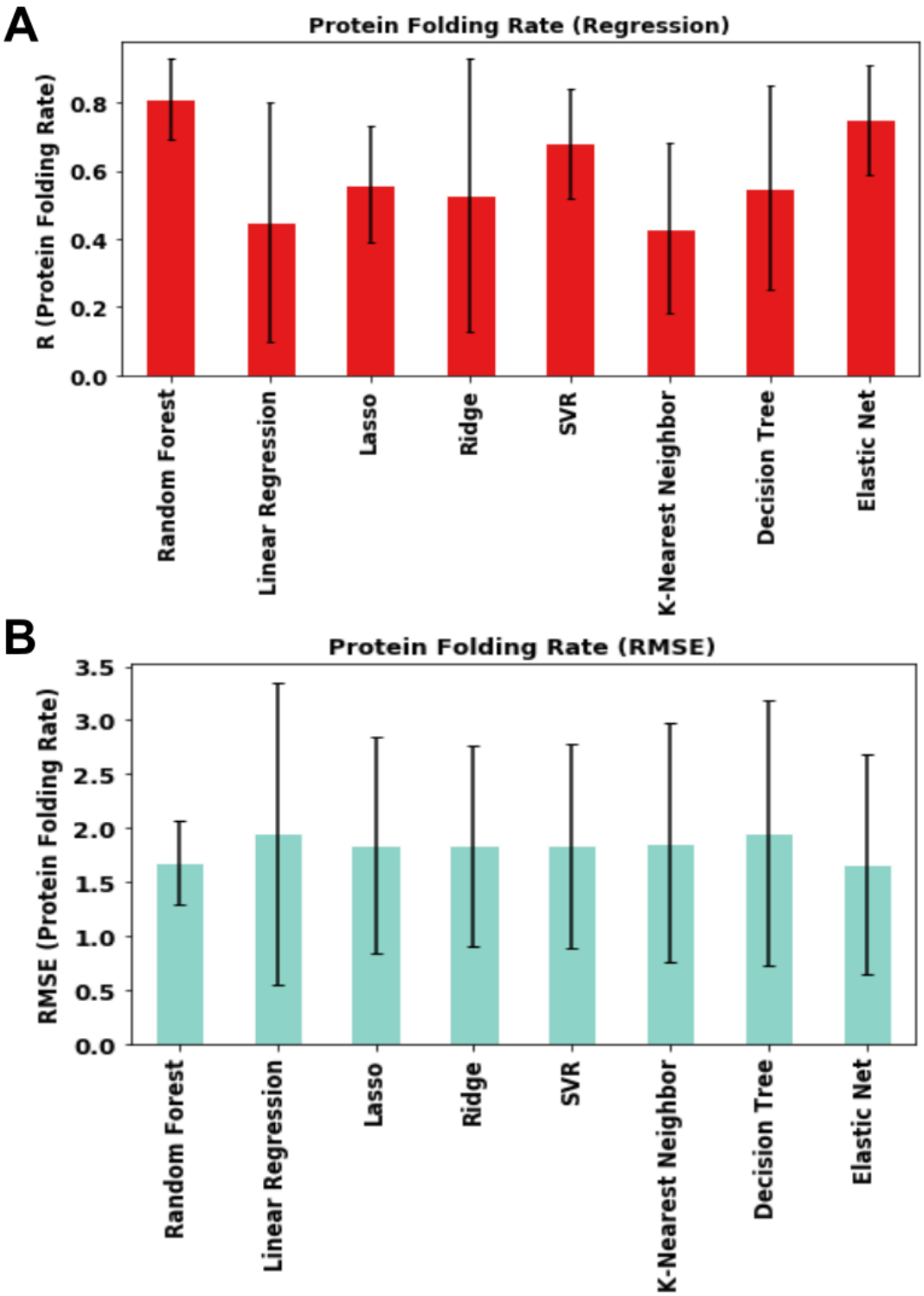

Figure 4.3: A) correlation value (R) for estimating the rate of protein folding is presented. The machine learning models are supplied with features taken from ProteinGW. Comparisons are made between ElasticNet, Decision Tree, Random Forest, KNN, SVR, Ridge, Lasso, and Linear Regression. (B) The plot shows models' Root mean squered error (RMSE) FOR protein folding rate prediction. (Mishra et. al. (249))

points, we successively reduced the number of training samples for modelling transmembrane/globular properties. The results are shown in Figure 4.4, where the Macro-F1 Score for ProteinGW ranges from 0.87 at nearly all fractions. However, for all quantities of training data, the performance of the graph-classifying CNN (232) is significantly worse. Additionally, when this proportion is marginally decreased, CNN's performance gradually deteriorates. As shown in Figure 4.4, we have also contrasted our approach with the graph Fourier transform-based filtering approach (231). The performance of the graph wavelet is consistently 92 percent in terms of accuracy even when just 50 percent of the training data is used.

We also compared all-α and all-β classification, and found that ProteinGW consistently performed well even when the number of training data points decreased (Figure 4.4). While CNN model accuracy decreased further between 0.66 to 0.59, ProteinGW accuracy decreased from 0.80 to 0.78 at the 50 percent of the training set. Additionally, GFT accuracy was 0.67 to 0.53.

The results of a more thorough evaluation for solubility are shown in Figure 4.4. As training data was trimmed from 85 to 50% of the total training data, the ROC AUC for ProteinGW changed from 0.73 to 0.66. For GFT, it varied between 0.63 and 0.59. Similar to CNN, the range was 0.61 to 0.58. Overall, such research shows that our strategy performs better at all fractions of the training set than other investigated methods.

4.2.3 Pattern and Importance of Graph Wavelet-based Feature Scores of Amino Acids in Determining Protein's Biophysical Property

The technique of multispectral decomposition method to analyse the protein residue interaction network with the signal on the node has rarely been used so far. Therefore, it is crucial to study the characteristics of features extracted and get insights of their pattern. For each protein function modelling, we looked at the pattern in the ranking of the scores based on the multispectral representation of the physicochemical and network characteristics of amino-acid residues. From the random forest machine learning model, we retrieved the significance of the graph wavelet-based feature score (GWFS) at four scales.

Gene Name	Mutation	Disease	References
LMNA	p.Arg25Pro (6YF5) (rs61578124)	Emery-Dreifuss muscular dystrophy 2, autosomal dominant (EDMD2) [MIM:181350]	https://www.uniprot.org/docs/humsavar https://www.uniprot.org/uniprot/P02545 https://onlinelibrary.wiley.com/doi/epdf/10.1002/ajmg.1463
CDKN2A	p.Met53Ile (2A5E) (rs104894095)	Melanoma, cutaneous malignant 2 (CMM2) [MIM:155601]	https://www.uniprot.org/docs/humsavar https://www.uniprot.org/uniprot/P42771
CA2	p.His94Tyr (1V9E) (rs id not available)	Osteopetrosis, autosomal recessive 3 (OPTB3) [MIM:259730]	https://www.uniprot.org/docs/humsavar
CDKN2A	p.Val95Ala (2A5E) (rs id not available)	Non-small cell lung carcinoma	https://www.uniprot.org/docs/humsavar https://www.uniprot.org/uniprot/P42771
CDKN2A	p.Glu119Gln (2A5E) (rs id not available)	A biliary tract tumor	https://www.uniprot.org/docs/humsavar https://www.uniprot.org/uniprot/P42771
EPHB4	p.Glu59Lys (2BBA) (rs1584667224)	Capillary malformation-arteriovenous malformation 2 (CMAVM2) [MIM:618196]	https://www.uniprot.org/docs/humsavar https://www.uniprot.org/uniprot/P54760
LMNA	p.Arg50Pro (6YF5) (rs60695352)	Emery-Dreifuss muscular dystrophy 2, autosomal dominant (EDMD2) [MIM:181350]	https://www.uniprot.org/docs/humsavar https://www.uniprot.org/uniprot/P02545 https://onlinelibrary.wiley.com/doi/epdf/10.1002/ajmg.1463
MET	p.His150Tyr (rs1436957498)	-	https://www.uniprot.org/docs/humsavar
ABO	p.Glu223Asp (rs id not available)	-	https://www.uniprot.org/docs/humsavar
ABO	p.Arg176Gly (rs387907096)	-	https://www.uniprot.org/docs/humsavar
EPHA3	p.Lys207Asn (rs200567888)	A pancreatic ductal adenocarcinoma sample	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2935892/ https://www.uniprot.org/docs/humsavar https://www.uniprot.org/uniprot/P29320

Table 4.2: Details of Protein mutation sites for case studies

Protein Folding Rate: Conservation score, node-weighted degree, refractive index, node degree, and residue frequency signals at scale 4 (low frequency) were the features most crucial for describing protein-folding rate, whereas hydrophobicity and polarity had little significance at scale 1 (high frequency). Similarly, scale 3 polarity and the conservation score both seemed to be somewhat significant (Figure 4.6A). Since low frequencies (high scales) in wavelet transform correspond to a signal's global information (or the concentration of many interactive residues with the same physiochemical property), it follows that a high folding rate occurs when many residues with a given property (such as node-degree or refractive index) are connected to one another. The folding rate is slower, however, if there are smaller clusters of these densely coupled residues or if the structure is made up of smaller modules.

Transmembrane-Globular: Additionally, we determined how crucial the characteristics of amino acids and how they are distributed on the residue-graph are for classifying globular and transmembrane proteins (see Figure 4.6C). Based on random forest-based classifiers, the most prominent characteristics were polarity and molecular weight at scales 1 and 4. This outcome may be explained by the fact that polar amino acids with lower interactivity are found at the surface of proteins with greater globularity (higher value of high-resolution signal at scale-1). Transmembrane proteins, in

contrast, contain nonpolar amino acids scattered throughout their surface such that polar amino acids interact with one another at the protein's core, resulting in a higher value for the low-frequency spectrum of polarity at scale-4 of polarity (Figure 4.6C). The coiled-coil tendency, whose low-frequency component seems to be significant, can be supported by a similar reasoning. It suggests that globular proteins may include bigger modules of interacting residues with a high propensity for coiling. Refractive index at wavelet scales 1 and 2 and turn tendency at scales 1 and 3 were other intriguing patterns in the significance of GWFS for modelling transmembrane/globular features.

α-β: The categorization of alpha and beta proteins (Figure 4.6B) also revealed that molecular weight, coil tendency, compressibility, and bulkiness at scale 4 were the most important factors. Alpha proteins would have bigger modules of interacting residue with a strong coil propensity since it is widely known that coiled-coil structures stabilise alpha-helices.

Solubility: However, our model indicated that scale 4 (poor resolution) refractive index seemed to be more significant than other feature scores for identifying soluble and insoluble proteins. Flexibility, node degree, residue frequency, and partial specific volume at scale 1 were other characteristics that had discernible value (Fig 4.6D). Overall, the significance of GWFS matching to a few residue characteristics and wavelet scales offers insights into the interactions between various kinds of amino acids to create various modules.

4.2.4 Graph Wavelet-based Features add Explainability About the Effect of Mutations on Protein's Biophysical Property

The benefit of our method is that it measures the multispectral magnitude of the score of significant and predictive characteristics for each amino acid. For instance, we may obtain the multispectral score based on graph of the most important predictive indicator (polarity) for each amino acid for the transmembrane/globular proteins classification. The significance of each amino acid in defining the type of protein—transmembrane or globular—can be explained using the feature score information for each amino acid. As a result, we looked at the effectiveness of our approach for determining a potential mechanism behind the impact of disease-associated mutations. With the use of the

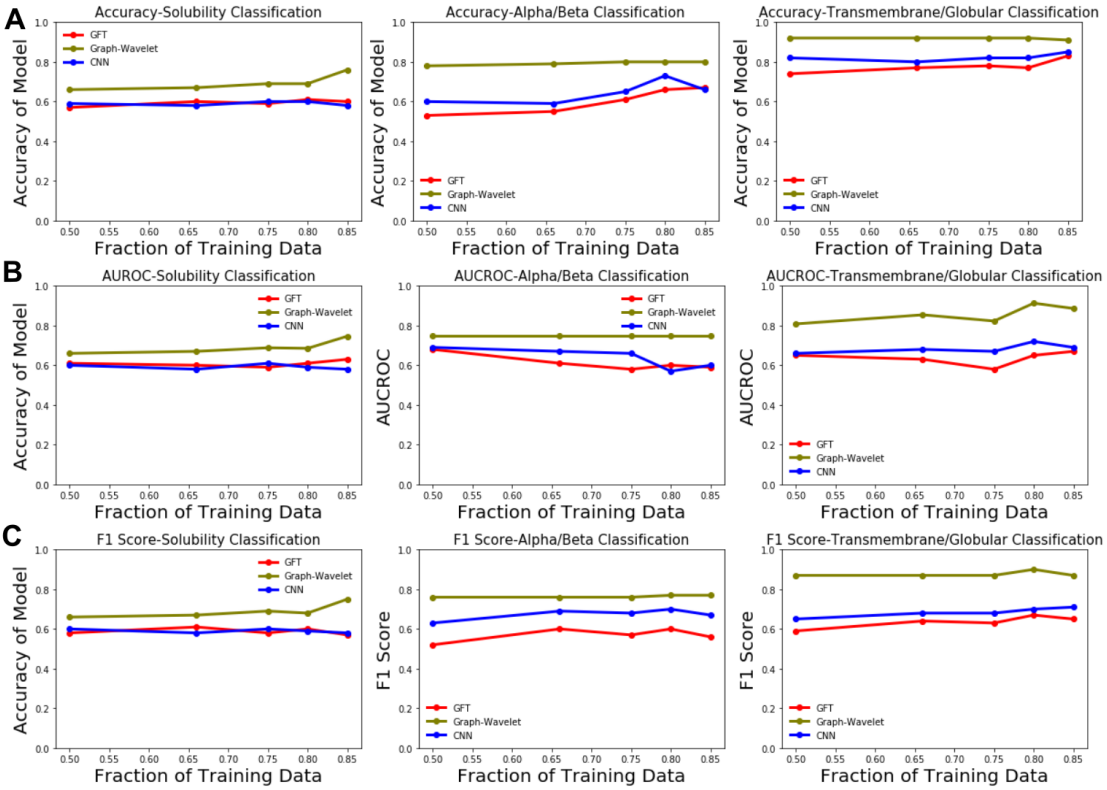

Figure 4.4: Comparison of ProteinGW with other methods and feature-weights for their amino-acid properties. (A) Accuracy of protienGW, convolutional neural network (CNN) and graph Fourier Transform (GFT) at the different number of training data points, The training fraction is reduced from 0.85 to 0.50. The importance of the feature score was calculated using a Random forest-based model for Solubility,Transmembbrane-Globular, All alpha-all beta (B) Similarly, AUROC is shown. (C) Here, F1-score is shown for all 3 properties. (Mishra et. al. (249))

UNIPROT (198) and HuVarBase (HUmanVARiantdataBASE) databases, we selected the disease-associated mutations in proteins utilized in our work for modelling protein globularity and folding rate (219). We were able to use our model to explain the potential effects of the mutations listed in Table 4.2. This explanation is described in more detail below.

Case studies based on Transmembrane/Globular property of Proteins

We investigated GWFS percentile scores at each amino acid's for the top features determined by feature-importance reported by random forest classifiers. Further, we evaluated at whether their mechanism of impact could be explained by high percentile scores GWFS for top predictive features for an amino acid for globularity prediction. For instance, a mutation in arginine to glycine/proline at mutation site 25 in one of

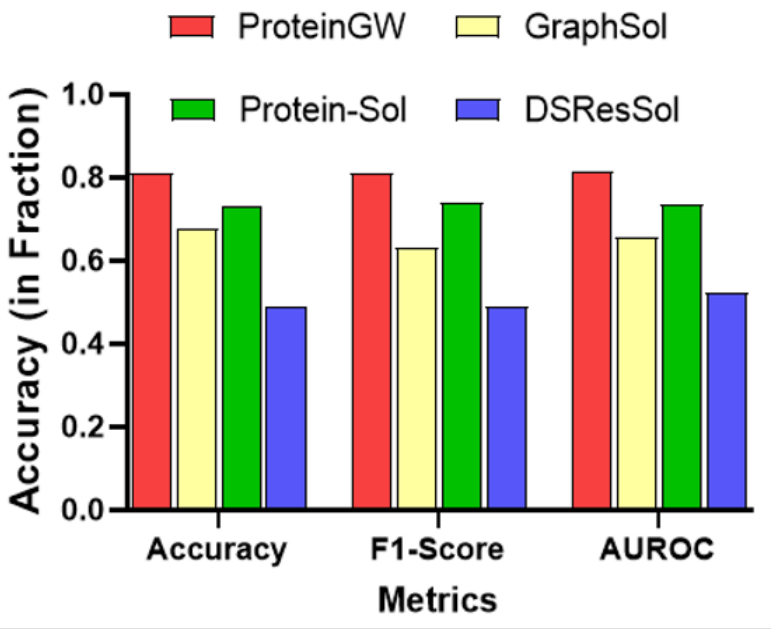

Figure 4.5: Benchmarking of Protein Solubility prediction methods.

the globular proteins Lamin A 17-70 coil1A dimer maintained by C-terminal capping (PDB ID:6YF5) has been linked to Emery-Dreifuss muscular dystrophy 2 (autosomal dominant form) (219) (EDMD2). Our preliminary study reveals that the corresponding mutation site has a comparatively higher than average polarity-based GWFS i.e nearly 75th percentile across all amino acids based on the average of all layers. In the PDB structure Lamin A (6YF5), we also discovered a second mutation at position 50 that is associated to EDMD2 (autosomal dominant). Additionally, the average GWFS for residue 50 in the LAMIN A structure (PDB id: 6YF5) is relatively high i.e 75% percentile (Figure 4.7A). Both of the residue sites (25 and 50) in 6YF5 had ranks below 72 percentile in the fourth layer of wavelet, which is crucial for modelling globularity, when we computed layer-wise GWFS for polarity (Figure 4.7).

multirow, makecell

The Human EphA3 Kinase Protein, which is most often mutated in lung cancer and other malignancies such as melanoma, glioma, and pancreatic (233) (234), as well as hepatocellular carcinoma and head and neck squamous cell carcinoma (233), was another site where we discovered high polarity based GWFS. The mutation in residue site 207 from leucine to asparagine is known to be related with the disorder pancreatic ductal adenocarcinoma, and it can be seen in the structure of the human EphA3 kinase domain in association with the inhibitor AWL-II-38.3 (PDB ID: 3DZQ) (Position:207). Among all amino acids, the polarity-based GWFS at position 207 is the

Method	Error	Error Statistic	Reference
Pred-PFR	2.03	RMS Error	Shen H-B, Song J-N, Chou K-C. Prediction of protein folding rates from primary sequence by fusing multiple sequential features, Journal of Biomedical Science and Engineering 2009.
FoldRate	2.03	RMS Error	Kuo-Chen C, Hong-Bin S. FoldRate: A web-server for predicting protein folding rates from primary sequence, The Open Bioinformatics Journal 2009;3.
SWFoldRate	2.27	Standard Error	Cheng X., Xiao X., Wu Zc et al. Swfoldrate:Predicting protein folding rates from amino acid sequence with sliding window method, Proteins: Structure, Function, and Bioinformatics 2013;81:140-148.
Graph Fourier based	1.92	RMS Error	Srivastava D, Bagler G, Kumar V. Graph Signal Processing on protein residue networks helps in studying its biophysical properties, bioRxiv 2021.
Graph wavelet based	1.685	RMS Error	Our Objective

Table 4.3: Benchmarking various Protein folding rate prediction methods.

highest (85.63). We also discovered another protein EPHB4 kinase domain inhibitor complex with a mutation (Glutamic acid to Lysine at position 59) known to be related with Capillary malformation-arteriovenous malformation within our collection of PDB structures used to mimic globularity (PDB ID: 2BBA). According to our analyses, the polarity of residue 59 in the structure of EPHB4 kinase has a high percentile score (81.422). Figure 4.7 displays the key top predictive feature percentile scores for the proteins mentioned above. the wavelet level-wise rank and average percentile of the polarity-based GWFS.

Using our trained model for the transmembrane/globular properties of proteins, we further examined the effects of mutations via a change in estimated globularity. Upon

substituting proline for arginine (at Position:25), the predicted likelihood of Lamin A structure (6YF5) being globular dropped only by 15% (from 0.71 to 0.60). (Figure 4.7). The predicted chance for the Lamin A structure (PDB ID: 6YF5) to be globular reduced from 0.79 to 0.58. We also calculated changes in expected likelihood for globularity after simulating mutations at 10 randomly selected residue sites on the Lamin A structure (6YF5). Ten random mutations had an average effect on globularity probability that was almost equivalent to replacing residues at positions 25 and 50. On the other hand, the projected likelihood after the mutation from glu to asp at position 223 in the transmembrane protein structure of Histo-blood group ABO system transferase (PDB Id:4Y63) was dropped by 40%. (from 0.99 to 0.55). The chance of transmembrane property decreased by around 25% as a result of 10 random mutations in the Histo-blood Group ABO System Transferase (PDB Id:4Y63) at the same time. Similar to this, after mutations at residues 207 and 59 in transmembrane proteins with PDB IDs 3DZQ(EphA3) and 2BBA(EPHB4), respectively, the chance of becoming trans-membrane decreased from 0.86, 0.96, to 0.41, 0.38. EphA3's mutation Lys207Asm significantly reduces transmembrane likelihood by more than 50%, supporting the theory that EphA3 may have lost its surface localisation, which has been linked to cancer formation (234). Only 25% of the protein structures 3DZQ and 3BBA were reduced by the simulated mutations at 10 random sites. The likelihood that the protein belonged to the transmembrane class before and after the mutation in Histo-blood Group ABO System Transferase (4Y63, Glu223Asp) is also depicted. Overall, our findings suggested that a plausible reason for the negative effects of matching mutations is the loss of transmembrane characteristic caused by mutation at a position with increased polarity based GWFS (Glu223Asp in Histo-blood Group ABO System Transferase, Lys207Asm EphA3 and Glu59Lys in EPHB4). However, for the LAMIN A structure, the GWFS of polarity at the corresponding residue were not that high and the reduction in globularity caused by the simulated mutations at residues 25 and 50 was not very significant, suggesting that the aforementioned mutations may have another mechanism of action (Arg25Pro and Arg50Pro).

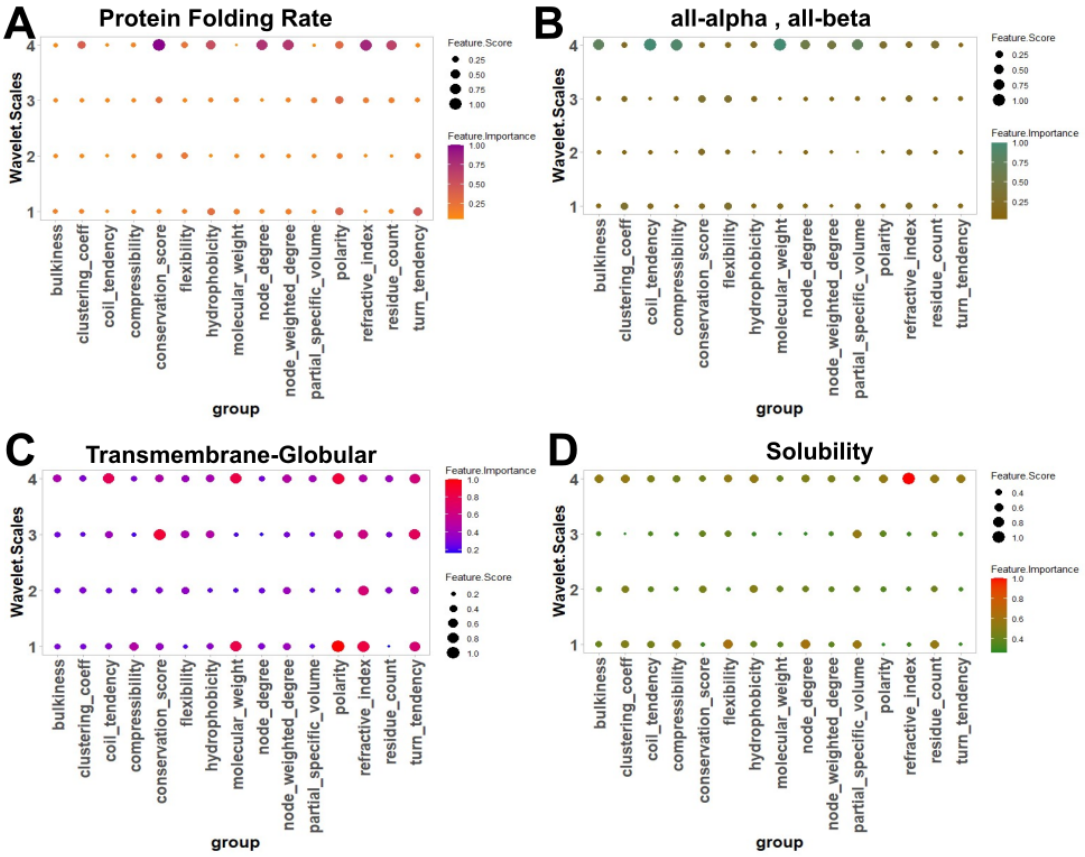

Figure 4.6: Graph wavelet-based features scores (GWFS) on various scales in simulating protein biophysical properties A) Protein-folding rate modelling with four wavelet scales scores is illustrated. Plot shows that at scale 4 (corresponding to low frequency), conservation score, node weighted degree, refractive index, node degree, and residue count are critical, but polarity is more significant at scale 1. Similar to this, polarity at scale 3 also seemed to be important. (B) 4 wavelet scales for distinguishing all-alpha and all-beta proteins are shown. The most relevant features were molecular weight, coil propensity, compressibility, and bulkiness at scale 4. In a similar way, (C) polarity, molecular weight at scales 1 and 4, conservation score at scale 3, coil tendency at scale 4, refractive index at scales 1 and 2, and turn tendency at scales 1 and 3 have shown to be more significant for transmembrane-globular. (D) Similarly, the refractive index at scale 4 and flexibility, node degree, residue frequency, and partial specific volume at scale 1 seemed to be more significant for the categorization of soluble and insoluble proteins. (Mishra et. al. (249))

Case Studies on the effect of mutations on Folding rate

For the proteins used in predicting protein folding rate, we additionally compiled known disease-associated mutations for those proteins. After aggregating their values at four levels of the wavelet spectrum, we looked at percentile scores for significant features in protein folding rate prediction. One such protein is the tumour suppressor P16INK4A

(PDB id: 2A5E), which has a mutation that links it to cutaneous malignant melanoma at position 53 (Met53Ile). A high percentile score for conservation (76.76) and a comparatively higher percentile score for flexibility can be noted for the corresponding residue position 53. (Figure 4.8). We also investigated two other mutations in the protein P16INK4A that have been associated to non-small cell lung cancer: glucine to glycine at residue position 119 (Glu119Gln) and valine to alanine at residue 95 (Val95Ala).

We took into consideration the crystal structure of bovine carbonic anhydrase II (CAII) (PDB id: 1V9E), as osteopetrosis is associated with a mutation in its human homologue at position 94 (histidine to tyrosine). Based on conservation, the refractive index, which emerged as a crucial component for folding rate, the relevant region in Bovine CAII protein has a marginally high GWFS (Figure 4.7A). We also utilised our pre-trained model to see whether the estimated protein folding rate changed as a result of the disease-causing mutations at the specified regions. Figure 4.7B displays the modifications in protein folding rate brought on by simulated mutations. Changes for 2A5E (Met53Ile) and 1V9E (His94Tyr) were 4.10809 and 12.11, respectively (526 percent change). The changes are 4.10527 and 4.10739 for 2A5E(Val95Ala) and 2A5E(Glu119Gln), respectively. The impact of mutations at random locations on the rate of protein folding in the protein structures 2A5E and 1E9E was not consistent with the research on deleterious mutations here. Even while we discovered that mutations significantly affected folding rate, this cannot be proved as the origin of the mutation's deleterious attribute. However, our approach has the ability to provide a hypothesis about how the deleterious mutation affect folding rate, which may aid researchers in developing relevant tests.

4.3 Discussion

Our research reveals a broadly applicable use of spectral graph wavelets for protein feature extraction. The possibility for these traits to be used to modelling different physical properties of proteins has been combined by ProteinGW. ProteinGW uses spectral graph wavelets to explain the understanding of the global and local significance of a feature for improved protein property prediction. To further filter the signal noise, our technique additionally determines an ideal threshold for each residue wavelet coefficient. We have

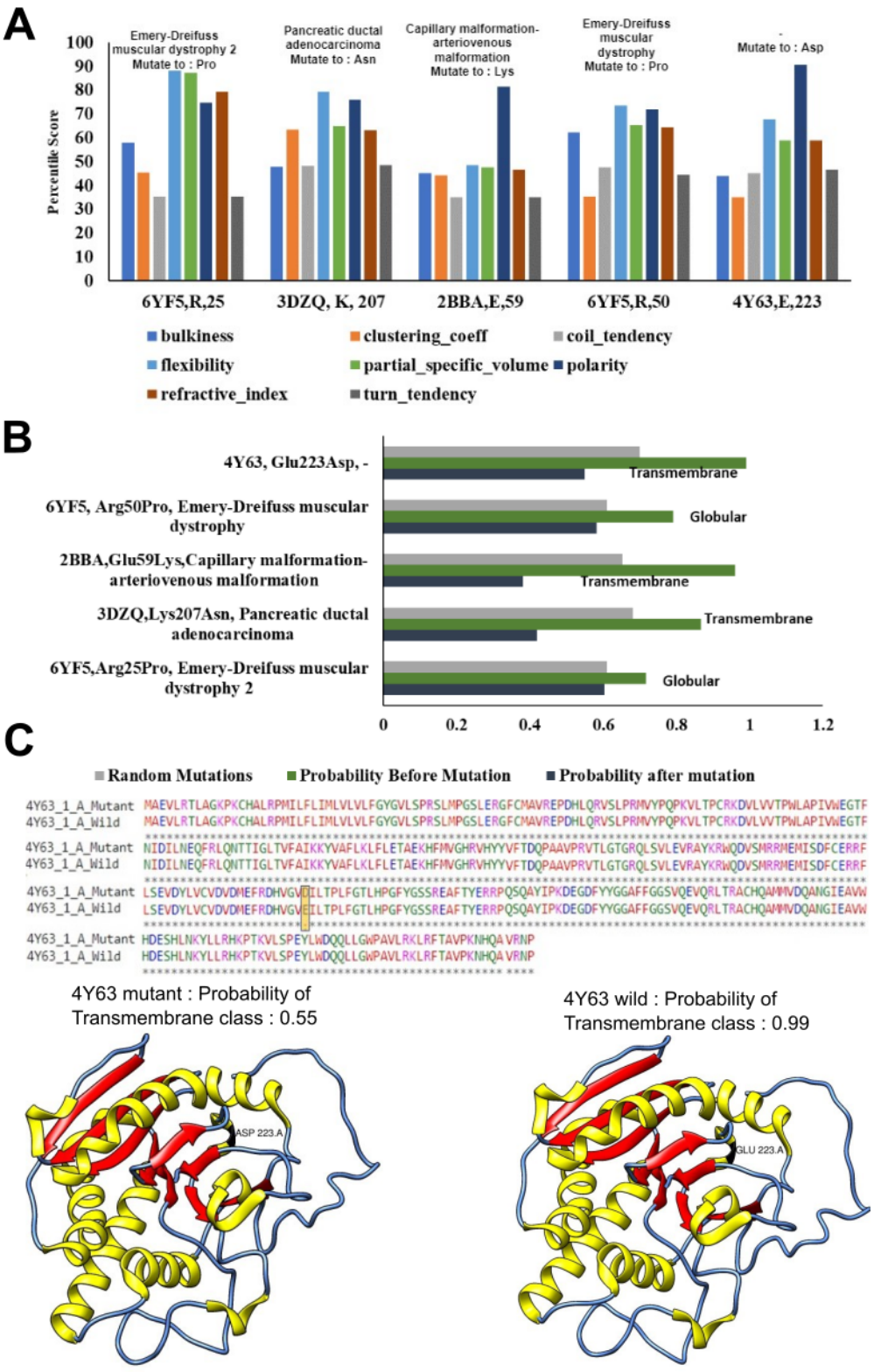

Figure 4.7: Transmembrane/globular property in effect of disease causing mutation(A)Percentile Scores of top features for proteins PDB:3DZQ(EPHA3),2BBA(EPHB4),6YF5(LMNA),6YF5(LMNA),4Y63(ABO gene).Average value for features across all four wavelet spectrum is shown.(B)Probability of proteins before,after mutation of original class.For control,average change in probability for mutations at 10 random sites is shown for every protein.(C)Visualization of 4Y63 mutation site. (Mishra et. al. (249))

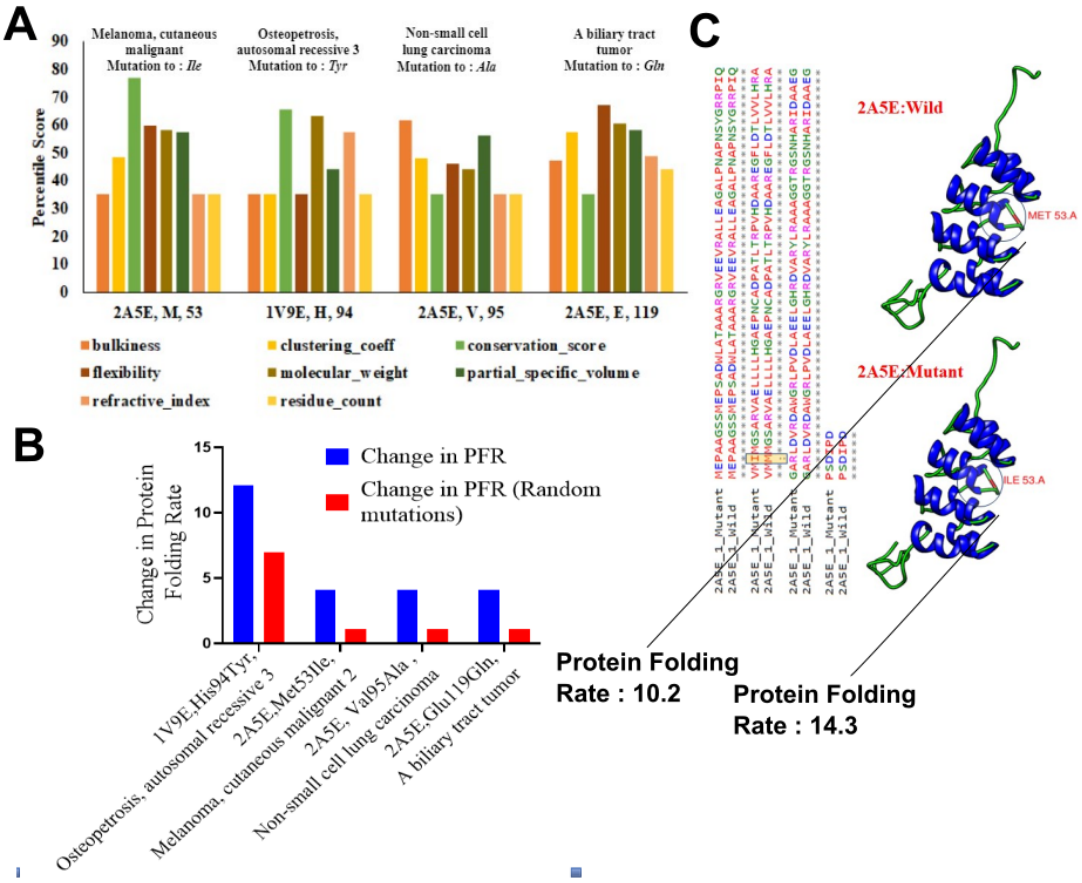

Figure 4.8: Assessing impact of disease-causing mutation on folding rate(A) Percentile Scores for top features for proteins 2A5E (Met53Ile, Melanoma), 1V9E (His94Tyr, Osteopetrosis), 2A5E (Val95Ala, Non-small cell lung carcinoma), and 2A5E(Glu119Gln, A biliary tract tumor) are bulkiness, conservation score, flexibility, refractive index, clustering coefficient, partial specific volume, residue frequency, and molarity . (B) The difference in rate of protein folding before and after a mutation. Every protein's average change in likelihood due to mutations at 10 random locations is also shown. (C) A visual representation of the 2A5E mutation site (Met53Ile, Melanoma, and Cutaneous Malignant) and the anticipated, actual, and changed protein folding rates. (Mishra et. al. (249))

shown the effectiveness of our technique for classifying transmembrane, globular, soluble, all-alpha, and all-beta proteins as well as for estimating the rate of protein folding. Nevertheless, every protein attribute may be modelled using the suggested technique. In order to predict four protein attributes, our technique has also exceeded CNN and feature extraction using the Graph Fourier transform. We also contrasted RMSE with three other approaches in order to assess the folding rate prediction.

The fact that the graph-based Fourier technique does not allow for evaluating the impact of each amino acid on the biophysical properties of proteins is a significant disadvantage. The graph-wavelet based technique, on the other hand, enables the in-

vestigation of each residue's impact and the prediction of its relationship with known deleterious consequences (disease-causing mutation). As a result, ProteinGW demonstrates explainability to wavelet-based predictive modelling, which helps users have a better understanding of the significance of each feature for each individual property. Given that the training set for ProteinGW only contains four attributes, it is not essential that it will always identify the potential mechanism of a mutation's impact. In such situation, it may be found in silico. The results of our analysis, for instance, point to potential connections between the proteins EPHA3 (PDB id: 3DZQ), LMNA (PDB id: 6YF5), EPHB4 (PDB id: 2BBA), and polarity score at residue sites 207, 50, and 59, respectively, which may be the cause of these proteins' well-known associations with various disorders. The effects of the simulated mutations at a single important residue site (Glu59Lys and Glu207Asm) were substantially greater, especially for the cell membrane proteins EPHB4 and EPHA3. These findings demonstrate how a little change in the signal of an amino acid's physiochemical properties, without altering the residue-interaction graph, may have a significant impact. The 3-dimensional modular organisation of physicochemical effects like hydrophobicity and the strategic arrangement of residues connecting various modules may be the root cause of such a broad impact. The community of related residues having the same physicochemical attribute signals is described in this module. Therefore, the pattern of linked components of the physicochemical signal at various resolutions, in addition to the structure of the residue-interaction graph, also determines the significance of a residue. Our approach can identify these patterns in the structure of proteins and utilise them to produce a hypothesis about the impact of exonic mutations that can then be used for a more thorough investigation.

In the field of predicting the impact of mutations on protein function and dynamics, various methods have been proposed. Sequence-based methods, such as SIFT/CADD or PolyPhen, primarily utilize protein sequences and their associated characteristics, such as the position-specific substitution matrix (PSSM). On the other hand, structure-based methods typically employ machine learning techniques that rely on training datasets. However, the influence of an individual amino acid on specific protein properties is often not well understood, which leads to skepticism among researchers regarding the reliability of predictive methods, particularly for new protein structures. To address this concern and enhance the confidence in predictive models, we have taken further steps

in our research.

In addition to predicting the effect of mutations, we have incorporated the estimation of the importance of multispectral features and the incorporation of explainability within machine learning models. By considering multiple spectral features, we aim to capture a broader range of information that could influence protein properties and better assess the impact of mutations. Furthermore, the inclusion of explainability techniques enables us to gain insights into the reasoning behind the predictions made by machine learning models, making the results more interpretable and transparent.

By integrating these additional steps into our approach, we strive to improve the reliability and trustworthiness of mutation effect predictions. We acknowledge the need for a comprehensive understanding of the factors affecting protein properties and seek to provide researchers with a more robust framework for predicting the effects of mutations on protein function and dynamics.

Our study of the residue interaction network using graph wavelet transform also provides a way to categorise proteins. The issue of structure-based categorization of proteins (SCOPE) has persisted, but in the post-genome-sequencing age, it's also critical to understand the impact of mutations. ProteinGW's ingenious use of the graph-wavelet has potential for fixing both challenges. Finally, the unsupervised classification of proteins may benefit from feature extraction from protein structures utilising graph-wavelet based on ProteinGW. As a result, ProteinGW may prove to be a valuable tool for many studies. ProteinGW has the drawback of relying on a training set to report the potential effects of mutations. Let's say a mutation affects a protein's biophysical characteristic for which there is a scarcity of training data. In such instance, adopting a graph-wavelet based strategy to emphasise the precise mechanism of its influence may not be simple. Therefore, in the future, we intend to accumulate additional training sets for many different protein features.

CHAPTER 5

Conclusion

Due to the diversity of illnesses, omics markers of disease are crucial for personalized therapy. Despite the development of computational tools, there are still just a few techniques that can capture the latent interactions between the many genes and amino acids that make up proteins. This gap may be filled by the graph-based learning technique, both supervised and unsupervised, which enables the creation of physiologically driven learning problems on graphs. Utilizing a collection of tools for processing graph signals—functions defined over the vertices in a graph—we used graph signal processing. Therefore, the goals we met in our thesis study are as follows.

5.1 Summary of Contribution

5.1.1 Denoising Large Read Count Matrices of Single-cell Expression Profiles with Graph Signal Processing for Better Network Inference

Unprecedented information on the effects of internal and external influences may be learned by using a regulatory-networks based technique for single-cell expression profiles. The accurate estimate of gene and regulatory connections might be hampered by noise and batch effects in sparse single-cell expression profiles. Here, we provide a theoretically novel technique for enhancing GWNet-based transcriptome analysis utilising graph-wavelet filters. Multiple gene-network inference techniques performed better owing to our approach. Most importantly, even in the presence of batch effect, GWNet enhanced consistency in gene-regulatory network prediction using single-cell transcriptome. The projected gene network's consistency allowed for accurate estimations of the changes in the effect of genes that were not specifically emphasised by the differential-expression study. When GWNet was applied to the single-cell transcriptomic profiles of lung cells, it was discovered that ageing had a physiologically significant impact on the

effect of pathways and master-regulators. Surprisingly, patterns caused by the effects of a new coronavirus infection in the human lung and the regulatory impact of age on pneumocyte type II cells exhibited a notable resemblance.

One advantage of our type of evaluation is that it may identify a few particular therapeutic targets that need more research. In clinical trials for pulmonary fibrosis, a kinase named JNK that binds and phosphorylates c-Jun is being examined. Treatment with androgen deprivation has shown to provide a limited degree of protection against SARS-COV-2 infection. In keeping with this tendency, our study suggests that *Etv5* may also be a therapeutic target to lessen the impact of age-related RAS pathway activation in the lung.

5.1.2 Graph based Integration of Large Single-cell Epigenome Profiles with Different Batch Effects

The proposed graph based co-embedding of scATACseq and scRNAseq demonstrated the advantages of mapping single-cell epigenomes to reference-cell profiles. Results based on knowledge-oriented reference-based edge weight learning surpassed integrative approaches like Seurat, LIGER, and Conos. It also revealed a pattern that showed superior embedding of open-chromatin profiles when distance was calculated based on projection on reference cell profiles for feature extraction and cell distance calculation than with SCALE, MINT, and SCANORAMA or other latent feature extraction technique. Moreover, when latent features are learnt using other methods, features in minority populations of cells that would have been masked by characteristics of cell-states in the majority might have been highlighted by clustering with reference cells.

Single-cell epigenome profiles are accurately and robustly mapped on large pool of reference expression profiles and matching cells are interpreted in a meaningful way using our proposed methodology. Technical and batch biases, such as cell-to-cell variability in signal i.e peak accessibility and noise, differing read-depths, methods, platforms, and labs, make this problem difficult to solve. Hence, in order to minimize batch and technical biases, our technique relies on median expression of top genes and peaks, rather than distance-based or latent feature extraction methods.

The devised method, which uses a graph-based approach, aids in discovering ac-

curate distances (edge-weights) between cells from varied species/batches with various peak lists. A unique method for calculating the distance between cells by projecting them upon a reference pool of datasets is included. The method relies on knowledge-based learning of edge weights between cells, which leads to a significant improvement in the joint embedding of cells. This helps to get rid of the batch effect and establish a reliable cell neighbourhood.

By jointly embedding query datasets on the basis of knowledge from reference profiles, this method produces useful patterns in disease profiles. The approach also shows applications like results pointing to varied levels and varieties of de-differentiated MPAL profiles. The flexibility and ability to flip lineages of MPAL profiles may be explained by their co-embedding with various progenitor cells.

As a result, reference-based edge-weight learning produces co-embedding and demonstrates other applications, including understanding the dedifferentiated states of cancer cells, highlighting imprints within individual cells that are a sign of apoptosis and a stress response in a subpopulation of embryonic stem cells, evaluating the enhancer landscape activity in single-cells, incorporating more single-cell histone modification datasets, and providing a powerful reference based search for query cells for key epigenetic regulators.

5.1.3 Explainable Predictive Model using Graph-Wavelet for Modeling Biophysical Properties of Proteins and Measuring Mutational Effects on Diseases

Integrative analysis techniques are required in the present post-genomic age to integrate data from protein structures with other relevant information to assess their biophysical characteristics and the consequences of non-synonymous mutations. Proteins include multispectral patterns of various physicochemical properties of amino acids, which may be used to understand how proteins behave. In order to describe the biophysical characteristics of amino acids, we here offer a technique based on the graph-wavelet transform of signals of features of amino acids in protein residue networks based on their structures. Additionally, it fared better than approaches based on convolutional neural networks and graph Fourier in predicting the physicochemical characteristics of pro-

teins. Our technique can quantify the impact of each amino acid on the physicochemical characteristics of proteins, even though it cannot forecast harmful mutations. Such an assessment of amino acid effects has the ability to elucidate the mechanism behind the detrimental non-synonymous mutation effect. Therefore, for better categorization and deeper comprehension, our technique may highlight patterns of distribution of amino-acid characteristics in the context of a biophysical feature in the structure of the protein.

5.2 Future Work

In order to improve denoising using GWNet, it is required to replace hard thresholding with automated system to find appropriate threshold for each dataset and there is a need to find an automated method to find value of "k" in KNN based graph which is initially built between cells.

However for scEpiSearch, future plans include expanding the use of single-cell histone modification datasets, evaluating enhancer landscape activity in single-cells, and providing an effective search engine for key epigenetic regulators.

ProteinGW has the drawback of relying on a training set to report the potential effects of mutations. Let's say a mutation affects a protein's biophysical characteristic for which there is a dearth of training data. In such instance, adopting a graph-wavelet based strategy to emphasise the precise mechanism of its influence may not be simple. Therefore, in the future, we intend to amass additional training sets for many different protein features.

REFERENCES

[1] Sandryhaila, Aliaksei, and José MF Moura. "Discrete signal processing on graphs." IEEE transactions on signal processing 61.7 (2013): 1644-1656.

[2] Estrada, E. (2012). The structure of complex networks: theory and applications. Oxford University Press.

[3] Rodrigues, Francisco Aparecido. "Network centrality: an introduction." A mathematical modeling approach from nonlinear dynamics to complex systems. Springer, Cham, 2019. 177-196.

[4] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine.In Proceedings of the Seventh International Conference on World Wide Web 7, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B.V

[5] Vince Grolmusz (2015). "A Note on the PageRank of Undirected Graphs". Information Processing Letters. 115 (6–8): 633–634. arXiv:1205.1960. doi:10.1016/j.ipl.2015.02.015.

[6] Chuang, H. Y., M. Hofree, et al. (2010). A decade of systems biology. Annu Rev Cell Dev Biol 26: 721–44

[7] Marbach, D., J. C. Costello, et al. (2012). Wisdom of crowds for robust gene network inference. Nat Methods 9(8): 796–804

[8] Tu Z, Wang L, Arbeitman MN et al. An integrative approach for causal gene identification and gene regulatory pathway inference, Bioinformatics 2006;22:e489-e496

[9] Iacono G, Massoni-Badosa R, Heyn H. Single-cell transcriptomics unveils gene regulatory network plasticity, Genome Biology 2019;20:110

[10] Pradhan A, Siwo GH, Singh N et al. Chemogenomic profiling of Plasmodium falciparum as a tool to aid antimalarial drug discovery, Scientific Reports 2015;5:15930.

[11] Maetschke SR, Madhamshettiwar PB, Davis MJ et al. Supervised, semi-supervised and unsupervised inference of gene regulatory networks, Briefings in Bioinformatics 2014;15:195-211.

[12] Margolin AA, Wang K, Lim WK et al. Reverse engineering cellular networks, Nature Protocols 2006;1:662-671.

[13] Skinnider MA, Squair JW, Foster LJ. Evaluating measures of association for single-cell transcriptomics, Nature Methods 2019;16:381-386.

[14] Sabatti, C., L. Rohlin, et al. (2002). Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res 30(13): 2886–93.

[15] Shi, Z., C. K. Derow, et al. (2010). Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. BMC Syst Biol 4: 74.

[16] Butte, A. J., and I. S. Kohane (1999). Unsupervised knowledge discovery in medical databases using relevance networks. Proc AMIA Symp: 711– 15.

[17] Margolin, A. A., I. Nemenman, et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1: S7

[18] Meyer, P. E., K. Kontos, et al. (2007). Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol. 79879

[19] Reshef, D. N., Y. A. Reshef, et al. (2011). Detecting novel associations in large data sets. Science 334(6062): 1518–24

[20] Milo, R., S. Shen-Orr, et al. (2002). Network motifs:simple building blocks of complex networks. Science 298(5594): 824–27.

[21] Alon, U. (2007). Network motifs: theory and experimental approaches. Nat Rev Genet 8(6): 450–61.

[22] Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. Cell 174, 1015–1030.e16 (2018).

[23] Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell 177, 1873–1887.e17 (2019).

[24] Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490 (2015).

[25] Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 348, 910–914 (2015).

[26] Nathan A, Baglaenko Y, Fonseka CY, et al. Multimodal single-cell approaches shed light on T cell heterogeneity. Curr Opin Immunol 2019;61:17–25.

[27] Bock C, Farlik M, Sheffield NC. Multi-Omics of single cells: strategies and applications. Trends Biotechnol 2016;34:605–8.

[28] Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet 2019;20:257–72.

[29] Leonavicius K, Nainys J, Kuciauskas D, et al. Multi-omics at single-cell resolution: comparison of experimental and data fusion approaches. Curr Opin Biotechnol 2019;55:159–66.

[30] Zhu C, Preissl S, Ren B. Single-cell multimodal omics: the power of many. Nat Methods 2020;17:11–4.

[31] Angermueller C, Clark SJ, Lee HJ, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods 2016;13:229–32.

[32] Clark SJ, Argelaguet R, Kapourani C-A, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun 2018;9:781.

[33] Argelaguet R, Velten B, Arnol D, et al. Multi-Omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol 2018;14:1–13.

[34] Argelaguet R, Arnol D, Bredikhin D, et al. MOFA+: a probabilistic framework for comprehensive integration of structured single-cell data. bioRxiv 2019;837104.

[35] Chen, S. and Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC bioinformatics, 19(1), 232.

[36] Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell 2019;177:1888–902.e21.

[37] Barkas N, Petukhov V, Nikolaeva D, et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. Nat Methods 2019;16:695–8.

[38] Pliner HA, Packer JS, McFaline-Figueroa JL, et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Mol Cell 2018;71:858–71.e8.

[39] Welch JD, Kozareva V, Ferreira A, et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell 2019;177:1873–87.e17.

[40] Campbell KR, Steif A, Laks E, et al. Clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. Genome Biol 2019;20:54.

[41] Jansen C, Ramirez RN, El-Ali NC, et al. Building gene regulatory networks from scATACseq and scRNA-seq using linked self organizing maps. PLoS Comput Biol 2019;15:e1006555.

[42] Welch JD, Hartemink AJ, Prins JF. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol 2017;18:138.

[43] Danese, Anna, et al. "EpiScanpy: integrated single-cell epigenomic analysis." Nature communications 12.1 (2021): 1-8.

[44] Polański K, Young MD, Miao Z, et al. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics 2020;36:964–5.

[45] Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with harmony. Nat Methods 2019;16:1289–96.

[46] Fang, Rongxin, et al. "SnapATAC: A comprehensive analysis package for single cell ATAC-seq." BioRxiv (2020): 615179.

[47] Yang Y, Li G, Qian H, et al. SMNN: batch effect correction for single-cell RNA-seq data via supervised mutual nearest neighbor detection. bioRxiv 2019;672261.

[48] Zhang F, Wu Y, Tian W. A novel approach to remove the batch effect of single-cell data. Cell Discov 2019;5:46.

[49] Nelson, D. L., Lehninger, A. L., and Cox, M. M. (2008). Lehninger principles of biochemistry. Macmillan.

[50] Westhead, D. R., Parish, J. H., and Twyman, R. M. (2002). Bioinformatics. Instant Notes. BIOS Sci. Pub. Ltd., Oxford, 257 pp.

[51] Bagler, G. and Sinha, S. (2007). Assortative mixing in protein contact networks and protein folding kinetics. Bioinformatics, 23(14):1760–1767.

[52] Chakrabarty, B. and Parekh, N. (2016). Naps: Network analysis of protein structures. Nucleic acids research, 44(W1):W375–W382.

[53] Yan, W., Zhou, J., Sun, M., Chen, J., Hu, G., and Shen, B. (2014). The construction of an amino acid network for understanding protein structure and function. Amino acids, 46(6):1419–1439.

[54] Szilágyi, A., Nussinov, R., and Csermely, P. (2013). Allo-network drugs: extension of the allosteric drug concept to protein-protein interaction and signaling networks. Current Topics in Medicinal Chemistry, 13(1):64–77.

[55] Jiang, X., Chen, C., and Xiao, Y. (2010). Improvements of network approach for analysis of the folding free-energy surface of peptides and proteins. Journal of computational chemistry, 31(13):2502–2509.

[56] Bagler, G. and Sinha, S. (2007). Assortative mixing in protein contact networks and protein folding kinetics. Bioinformatics, 23(14):1760–1767.

[57] Gromiha, M. M., Thangakani, A. M., and Selvaraj, S. (2006). Fold-rate: prediction of protein folding rates from amino acid sequence. Nucleic acids research, 34(suppl 2):W70–W74.

[58] Srivastava, D. and Kumar, V. (2018). Graph signal processing based analysis of biological networks. PhD thesis, IIIT-D.

[59] J.A. Capra, M. Singh, Predicting functionally important residues from sequence conservation, Bioinformatics 23 (2007) 1875–1882.

[60] Brouwer, A. E. and Haemers, W. H. (2011). Spectra of graphs. Springer Science Business Media.

[61] F. Pazos, M. Helmer-Citterich, G. Ausiello, A. Valencia, Correlated mutations contain information about protein–protein interaction, J. Mol. Biol. 271 (1997) 511–523.

[62] R.K. Kuipers, H.J. Joosten, E. Verwiel, S. Paans, J. Akerboom, J. van der Oost, N.G. Leferink, W.J. van Berkel, G. Vriend, P.J. Schaap, Correlated mutation analyses on super-family alignments reveal functionally important residues, Proteins 76 (2009) 608–616.

[63] C. Marino Buslje, E. Teppa, T. Di Domenico, J.M. Delfino, M. Nielsen, Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification, PLoS Comput. Biol. 6 (2010) e1000978

[64] U. Gobel, C. Sander, R. Schneider, A. Valencia, Correlated mutations and residue contacts in proteins, Proteins 18 (1994) 309–317.

[65] E.A. Ortlund, J.T. Bridgham, M.R. Redinbo, J.W. Thornton, Crystal structure of an ancient protein: evolution by conformational epistasis, Science 317 (2007) 1544–1548.

[66] DePristo MA, Weinreich DM, Hartl DL. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet. 2005;6(9):678–87. Pmid:16074985

[67] Povolotskaya IS, Kondrashov FA. Sequence space and the ongoing expansion of the protein universe. Nature. 2010;465(7300):922–6. Pmid:20485343

[68] Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005;353(2):459–73. Pmid:16169011

[69] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9. Pmid:20354512

[70] Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011;12(9):628–40. Pmid:21850043

[71] Wu J, Jiang R. Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. ScientificWorldJournal. 2013;2013:675851. Pmid:23431257

[72] Studer RA, Dessailly BH, Orengo CA. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem J. 2013;449(3):581–94. Pmid:23301657

[73] Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics. 2013;14 Suppl 3:S7. Pmid:23819521

[74] Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, et al. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol. 2014;10(1):e1003440. Pmid:24453961

[75] Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, et al. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet. 2012;91(6):1022–32. Pmid:23217326

[76] Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, et al. The functional spectrum of low-frequency coding variation. Genome Biol. 2011;12(9):R84. Pmid:21917140

[77] Lehner B. Genotype to phenotype: lessons from model organisms for human genetics. Nat Rev Genet. 2013;14(3):168–78. Pmid:23358379

[78] Gray VE, Kukurba KR, Kumar S. Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. Bioinformatics. 2012;28(16):2093–6. Pmid:22685075

[79] Burga A, Lehner B. Predicting phenotypic variation from genotypes, phenotypes and a combination of the two. Curr Opin Biotechnol. 2013;24(4):803–9. Pmid:23540420

[80] Hecht M, Bromberg Y, Rost B. News from the protein mutability landscape. J Mol Biol. 2013;425(21):3937–48. Pmid:23896297

[81] Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11(8):801–7. Pmid:25075907

[82] Humphris-Narayanan E, Akiva E, Varela R, S OC, Kortemme T. Prediction of mutational tolerance in HIV-1 protease and reverse transcriptase using flexible backbone protein design. PLoS Comput Biol. 2012;8(8):e1002639. pmid:22927804

[83] Tu Z, Wang L, Arbeitman MN et al. An integrative approach for causal gene identification and gene regulatory pathway inference, Bioinformatics 2006;22:e489-e496

[84] Iacono G, Massoni-Badosa R, Heyn H. Single-cell transcriptomics unveils gene regulatory network plasticity, Genome Biology 2019;20:110

[85] Pradhan A, Siwo GH, Singh N et al. Chemogenomic profiling of Plasmodium falciparum as a tool to aid antimalarial drug discovery, Scientific Reports 2015;5:15930.

[86] Maetschke SR, Madhamshettiwar PB, Davis MJ et al. Supervised, semi-supervised and unsupervised inference of gene regulatory networks, Briefings in Bioinformatics 2014;15:195-211.

[87] Margolin AA, Wang K, Lim WK et al. Reverse engineering cellular networks, Nature Protocols 2006;1:662-671.

[88] Skinnider MA, Squair JW, Foster LJ. Evaluating measures of association for single-cell transcriptomics, Nature Methods 2019;16:381-386.

[89] Langfelder, Peter, and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis." BMC bioinformatics 9.1 (2008): 1-13.

[90] Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data, BMC bioinformatics 2018;19:232.

[91] Aibar S, González-Blas CB, Moerman T et al. SCENIC: single-cell regulatory network inference and clustering, Nature Methods 2017;14:1083-1086.

[92] Matsumoto H, Kiryu H, Furusawa C et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation, Bioinformatics (Oxford, England) 2017;33:2314-2321.

[93] Chan TE, Stumpf MPH, Babtie AC. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures, Cell Systems 2017;5:251-267.e253.

[94] Kim JK, Kolodziejczyk AA, Ilicic T et al. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression, Nature communications 2015; 6(1):1-9

[95] Raser JM, O'Shea EK. Noise in gene expression: origins, consequences, and control, Science (New York, N.Y.) 2005;309:2010-2013.

[96] Lichtblau Y, Zimmermann K, Haldemann B et al. Comparative assessment of differential network analysis methods, Briefings in Bioinformatics 2017;18:837-850.

[97] Kimmel JC, Penland L, Rubinstein ND et al. Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging, Genome Research 2019;29:2088-2103.

[98] Zhou, Q. et al. (2007). A gene regulatory network in mouse embryonic stem cells. Proceedings of the National Academy of Sciences, 104(42),16438–16443. Publisher: National Academy of Sciences Section:Physical Sciences.

[99] Marbach D, Costello JC, Küffner R et al. Wisdom of crowds for robust gene network inference, Nature Methods 2012;9:796-804.

[100] Schaffter T, Marbach D, Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods, Bioinformatics (Oxford, England) 2011;27:2263-2270.

[101] Burkhardt DB, Stanley JS, Perdigoto AL, et al. Enhancing experimental signals in single-cell RNA-sequencing data using graph signal processing. bioRxiv. 2019: 10:532846.

[102] Ziegenhain C, Vieth B, Parekh S et al. Comparative Analysis of Single-Cell RNA Sequencing Methods, Molecular Cell 2017;65:631-643.e634.

[103] Zhou Q, Chipperfield H, Melton DA et al. A gene regulatory network in mouse embryonic stem cells, Proceedings of the National Academy of Sciences 2007;104:16438-16443

[104] van Dijk D, Sharma R, Nainys J et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion, Cell 2018;174:716-729.e727.

[105] Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nature communications. 2018:9(1):1-9.

[106] Eraslan G, Simon LM, Mircea M, et al. Single-cell RNA-seq denoising using a deep count autoencoder. Nature communications. 2019:10(1):1-4.

[107] Huang M, Wang J, Torre E, Dueck H, et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nature methods. 2018; 15(7):539-542.

[108] Aparicio L, Bordyuh M, Blumberg AJ, Rabadan R. A random matrix theory approach to denoise single-cell data. Patterns. 2020:4:100035.

[109] Troyanskaya O, Cantor M, Sherlock G et al. Missing value estimation methods for DNA microarrays, Bioinformatics (Oxford, England) 2001;17:520-525.

[110] Wagner, F. et al. (2018). K-nearest neighbor smoothing for high throughput single-cell RNA-Seq data. bioRxiv, page 217737. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[111] Enge M, Arda HE, Mignardi M et al. Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns, Cell 2017;171:321-330.e314.

[112] Chen EY, Tan CM, Kou Y et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC bioinformatics 2013;14:128.

[113] Chen M, Brown LA. Histamine stimulation of surfactant secretion from rat type II pneumocytes, The American Journal of Physiology 1990;258:L195-200.

[114] Lecce L, Lam YT, Lindsay LA et al. Aging impairs VEGF-mediated, androgen-dependent regulation of angiogenesis, Molecular Endocrinology (Baltimore, Md.) 2014;28:1487-1501.

[115] Al-Saiedy M, Pratt R, Lai P et al. Dysfunction of pulmonary surfactant mediated by phospholipid oxidation is cholesterol-dependent, Biochimica Et Biophysica Acta. General Subjects 2018;1862:1040-1049.

[116] Schouten LRA, Helmerhorst HJF, Wagenaar GTM et al. Age-Dependent Changes in the Pulmonary Renin-Angiotensin System Are Associated With Severity of Lung Injury in a Model of Acute Lung Injury in Rats, Critical Care Medicine 2016;44:e1226-e1235.

[117] Park S-K, Dahmer MK, Quasney MW. MAPK and JAK-STAT signaling pathways are involved in the oxidative stress-induced decrease in expression of surfactant protein genes, Cellular Physiology and Biochemistry: International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology 2012;30:334-346.

[118] Zhang Z, Newton K, Kummerfeld SK et al. Transcription factor Etv5 is essential for the maintenance of alveolar type II cells, Proceedings of the National Academy of Sciences of the United States of America 2017;114:3903-3908.

[119] Reddy NM, Vegiraju S, Irving A et al. Targeted deletion of Jun/AP-1 in alveolar epithelial cells causes progressive emphysema and worsens cigarette smoke-induced lung inflammation, The American Journal of Pathology 2012;180:562-574.

[120] Mikkonen L, Pihlajamaa P, Sahu B et al. Androgen receptor and androgen-dependent gene expression in lung, Molecular and Cellular Endocrinology 2010;317:14-24.

[121] Viola A, Munari F, Sánchez-Rodríguez R et al. The Metabolic Signature of Macrophage Responses, Frontiers in Immunology 2019;10:1462.

[122] Blanco-Melo D, Nilsson-Payant BE, Liu W-C et al. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19, Cell 2020;181:1036-1045.e1039.

[123] Qi F, Qian S, Zhang S et al. Single cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses, Biochemical and biophysical research communications 2020;526:135-140.

[124] Chow RD, Chen S. The aging transcriptome and cellular landscape of the human lung in relation to SARS-CoV-2, bioRxiv 2020:2020.2004.2007.030684

[125] Yew-Booth L, Birrell MA, Lau MS et al. JAK-STAT pathway activation in COPD, The European Respiratory Journal 2015;46:843-845.

[126] Sharifi N, Ryan CJ. Androgen hazards with COVID-19, Endocrine-Related Cancer 2020;27:E1-E3.

[127] Bryce PJ, Mathias CB, Harrison KL et al. The H1 histamine receptor regulates allergic lung responses, Journal of Clinical Investigation 2006;116:1624-1632

[128] Greenberg S, Horan G, Bennett B et al. Late Breaking Abstract - Evaluation of the JNK inhibitor, CC-90001, in a phase 1b pulmonary fibrosis trial, European Respiratory Journal 2017;50.

[129] Montopoli M, Zumerle S, Vettor R et al. Androgen-deprivation therapies for prostate cancer and risk of infection by SARS-CoV-2: a population-based study (N = 4532), Annals of Oncology: Official Journal of the European Society for Medical Oncology 2020.

[130] Maaten LVD and Geoffrey H, Visualizing data using t-SNE. Journal of machine learning research 2008; 9:2579-2605.

[131] Sandryhaila A, Moura JMF. Discrete Signal Processing on Graphs: Frequency Analysis, IEEE Transactions on Signal Processing 2014;62:3042-3054.

[132] F. K. Chung, Spectral Graph Theory, Vol. 92 of CBMS Regional Conference Series in Mathematics, AMS Bookstore, 1997.

[133] M. Reed, B. Simon, Methods of Modern Mathematical Physics Volume 1 : Functional Analysis, Academic Press, 1980.

[134] Priya, K. Devi, G. Sasibhushana Rao, and PSV Subba Rao. "Comparative analysis of wavelet thresholding techniques with wavelet-wiener filter on ECG signal." Procedia Computer Science 87 (2016): 178-183.

[135] D. L. Donoho, 1991. “De-noising by soft thresholding”, IEEE Transaction on Information Theory, Vol. 41, pp. 613–627, May 1995.

[136] Donoho D.L., Johnstone I.M., “Adapting to unknown smoothness via wavelet shrinkage” , J. Am. Statis. Ass.L.1995 .

[137] S. Grace Chang, Bin Yu, Martin Vetterli, 2000, “Adaptive Wavelet Thresholding for Image Denoising and Compression”, IEEE transactions on image processing, vol. 9, pp. 1532-1546.

[138] Javier Portilla, Vasily Strela, Martin J. Wainwright, and Eero P. Simoncelli, 2003, “Image Denoising Using Scale Mixtures of Gaussians in the Wavelet Domain”, IEEE transactions on image processing, vol. 12, pp. 1338-1351.

[139] Hammond DK, Vandergheynst P, Gribonval R. Wavelets on graphs via spectral graph theory, Applied and Computational Harmonic Analysis 2011;30:129-150

[140] Erb I, Notredame C. How should we measure proportionality on relative gene expression data?, Theory in Biosciences = Theorie in Den Biowissenschaften 2016;135:21-36.

[141] Quinn TP, Richardson MF, Lovell D et al. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis, Scientific Reports 2017;7:16252.

[142] Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks, Nucleic Acids Research 2017;45:D408-D414.

[143] Angelidis I, Simon LM, Fernandez IE et al. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics, Nature Communications 2019;10:963.

[144] Urban, Elizabeth A., and Robert J. Johnston Jr. "Buffering and amplifying transcriptional noise during cell fate specification." Frontiers in genetics 9 (2018): 591.

[145] Lee, J.T.H., Patikas, N., Kiselev, V.Y. and Hemberg, M. (2021) Fast searches of large collections of single-cell data using scfind. Nature Methods, 18, 262-271.

[146] Wu, K.E., Yost, K.E., Chang, H.Y. and Zou, J. (2021) BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. Proceedings of the National Academy of Sciences, 118.

[147] Wang, C., Sun, D., Huang, X., Wan, C., Li, Z., Han, Y., Qin, Q., Fan, J., Qiu, X. and Xie, Y. (2020) Integrative analyses of single-cell transcriptome and regulome using MAESTRO. Genome biology, 21, 1-28.

[148] Jin, S., Zhang, L. and Nie, Q. (2020) scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. Genome biology, 21, 1-19.

[149] Danese, A., Richter, M.L., Chaichoompu, K., Fischer, D.S., Theis, F.J. and Colome-Tatche, M. (2021) EpiScanpy: integrated single-cell epigenomic analysis. Nat Commun, 12, 5228.

[150] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P. and Satija, R. (2019) Comprehensive Integration of Single-Cell Data. Cell, 177, 1888-1902 e1821.

[151] Liu, J., Gao, C., Sodicoff, J., Kozareva, V., Macosko, E.Z. and Welch, J.D. (2020) Jointly defining cell types from multiple single-cell datasets using LIGER. Nature protocols, 15, 3632-3662.

[152] Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K. and Kharchenko, P.V. (2019) Joint analysis of heterogeneous single-cell RNA-seq dataset collections. Nat Methods, 16, 695-698.

[153] Luecken, M.D., Buttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colome-Tatche, M. et al. (2022) Benchmarking atlas-level data integration in single-cell genomics. Nat Methods, 19, 41-50.

[154] Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S. et al. (2018) A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell, 174, 1309-1324 e1318.

[155] Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y. and Greenleaf, W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. Nature, 523, 486-490.

[156] Corces, M.R., Shcherbina, A., Kundu, S., Gloudemans, M.J., Frésard, L., Granja, J.M., Louie, B.H., Eulalio, T., Shams, S. and Bagdatli, S.T. (2020) Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. Nature genetics, 52, 1158-1168.

[157] Lahnemann, D., Koster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A. et al. (2020) Eleven grand challenges in single-cell data science. Genome Biol, 21, 31.

[158] Jin, W., Tang, Q., Wan, M., Cui, K., Zhang, Y., Ren, G., Ni, B., Sklar, J., Przytycka, T.M., Childs, R. et al. (2015) Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. Nature, 528, 142-146.

[159] Lai, B., Gao, W., Cui, K., Xie, W., Tang, Q., Jin, W., Hu, G., Ni, B. and Zhao, K. (2018) Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. Nature, 562, 281-285.

[160] Cole, M.B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S. and Yosef, N. (2019) Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. Cell Syst, 8, 315-328 e318.

[161] Schult, D.A. and Swart, P. (2008), Proceedings of the 7th Python in science conferences (SciPy 2008). Pasadena, CA, Vol. 2008, pp. 11-16.

[162] McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M. and Bejerano, G. (2010) GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology, 28, 495-501.

[163] Chawla, S., Samydurai, S., Kong, S.L., Wu, Z., Wang, Z., Tam, W.L., Sengupta, D. and Kumar, V. (2021) UniPath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles. Nucleic acids research, 49, e13-e13.

[164] Fruchterman, T.M. and Reingold, E.M. (1991) Graph drawing by force-directed placement. Software: Practice and experience, 21, 1129-1164.

[165] Schult, D.A. and Swart, P. (2008), Proceedings of the 7th Python in science conferences (SciPy 2008). Pasadena, CA, Vol. 2008, pp. 11-16.

[166] Xiang, T. and Gong, S. (2008) Spectral clustering with eigenvector selection. Pattern Recognition, 41, 1012-1029.

[167] Xiong, L., et al., SCALE method for single-cell ATAC-seq analysis via latent feature extraction. Nat Commun, 2019. 10(1): p. 4576.

[168] Lopez, R., et al., Deep generative modeling for single-cell transcriptomics. Nat Methods, 2018. 15(12): p. 1053-1058.

[169] Hie, B., B. Bryson, and B. Berger, Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol, 2019. 37(6): p. 685-691.

[170] Rohart, F., et al., MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. BMC Bioinformatics, 2017. 18(1): p. 128.

[171] Luecken, M.D., Buttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colome-Tatche, M. et al. (2022) Benchmarking atlas-level data integration in single-cell genomics. Nat Methods, 19, 41-50.

[172] Fu, S., Wang, Q., Moore, J.E., Purcaro, M.J., Pratt, H.E., Fan, K., Gu, C., Jiang, C., Zhu, R. and Kundaje, A. (2018) Differential analysis of chromatin accessibility and histone modifications for predicting mouse developmental enhancers. Nucleic acids research, 46, 11184-11201.

[173] Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57-74.

[174] Bujold, D., Morais, D.A.L., Gauthier, C., Cote, C., Caron, M., Kwan, T., Chen, K.C., Laperle, J., Markovits, A.N., Pastinen, T. et al. (2016) The International Human Epigenome Consortium Data Portal. Cell Syst, 3, 496-499 e492.

[175] Srivastava, Divyanshu, et al. "CellAtlasSearch: a scalable search engine for single cells." Nucleic acids research 46.W1 (2018): W141-W147.

[176] Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F. et al. (2018) Mapping the Mouse Cell Atlas by Microwell-Seq. Cell, 173, 1307.

[177] Tetteroo, P.A., Massaro, F., Mulder, A., Schreuder-van Gelder, R. and von dem Borne, A.E. (1984) Megakaryoblastic differentiation of proerythroblastic K562 cell-line cells. Leuk Res, 8, 197-206.

[178] Jacoby, E., Nguyen, S.M., Fountaine, T.J., Welp, K., Gryder, B., Qin, H., Yang, Y., Chien, C.D., Seif, A.E. and Lei, H. (2016) CD19 CAR immune pressure induces B-precursor acute lymphoblastic leukaemia lineage switch exposing inherent leukaemic plasticity. Nature communications, 7, 1-10.

[179] Slany, R.K. (2009) The molecular biology of mixed lineage leukemia. Haematologica, 94, 984.

[180] Gallagher, R., Collins, S., Trujillo, J., McCredie, K., Ahearn, M., Tsai, S., Metzgar, R., Aulakh, G., Ting, R., Ruscetti, F. et al. (1979) Characterization of the continuous, differentiating myeloid cell line (HL-60) from a patient with acute promyelocytic leukemia. Blood, 54, 713-733.

[181] Imaizumi, M., Uozumi, J. and Breitman, T.R. (1987) Retinoic acid-induced monocytic differentiation of HL60/MRI, a cell line derived from a transplantable HL60 tumor. Cancer research, 47, 1434-1440.

[182] Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A. et al. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res, 44, W90-97.

[183] Sutherland, J.A., Turner, A.R., Mannoni, P., McGann, L.E. and Turc, J.M. (1986) Differentiation of K562 leukemia cells along erythroid, macrophage, and megakaryocyte lineages. J Biol Response Mod, 5, 250-262.

[184] Sarna, M.K., Ingley, E., Busfield, S.J., Cull, V.S., Lepere, W., McCarthy, D.J., Wright, M.J., Palmer, G.A., Chappell, D., Sayer, M.S. et al. (2003) Differential regulation of SOCS genes in normal and transformed erythroid cells. Oncogene, 22, 3221-3230.

[185] Fugazza, C., Barbarani, G., Elangovan, S., Marini, M.G., Giolitto, S., Font-Monclus, I., Marongiu, M.F., Manunza, L., Strouboulis, J. and Cantù, C. (2021)

The Coup-TFII orphan nuclear receptor is an activator of the -globin gene. haematologica, 106, 474.

[186] Zhao, C., Wang, B., Meng, D., Cao, Y., Yang, J., Zhao, X. and Chen, B. (2005) Study on rapid generation of dendritic cells from K562 cell line induced by A23187 alone. Zhonghua xue ye xue za zhi= Zhonghua Xueyexue Zazhi, 26, 408-412.

[187] Hie, B., Bryson, B. and Berger, B. (2019) Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol, 37, 685-691.

[188] Rohart, F., Eslami, A., Matigian, N., Bougeard, S. and Le Cao, K.A. (2017) MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. BMC Bioinformatics, 18, 128.

[189] Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. and Yosef, N. (2018) Deep generative modeling for single-cell transcriptomics. Nat Methods, 15, 1053-1058.

[190] Xiong, L., Xu, K., Tian, K., Shao, Y., Tang, L., Gao, G., Zhang, M., Jiang, T. and Zhang, Q.C. (2019) SCALE method for single-cell ATAC-seq analysis via latent feature extraction. Nat Commun, 10, 4576.

[191] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." kdd. Vol. 96. No. 34. 1996.

[192] Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y. et al. (2019) Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat Biotechnol, 37, 1458-1465.

[193] Kleiveland, C.R. (2015) In Verhoeckx, K., Cotter, P., Lopez-Exposito, I., Kleiveland, C., Lea, T., Mackie, A., Requena, T., Swiatecka, D. and Wichers, H. (eds.), The Impact of Food Bioactives on Health: in vitro and ex vivo models, Cham (CH), pp. 161-167.

[194] Buenrostro, J.D., Corces, M.R., Lareau, C.A., Wu, B., Schep, A.N., Aryee, M.J., Majeti, R., Chang, H.Y. and Greenleaf, W.J. (2018) Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. Cell, 173, 1535-1548. e1516.

[195] Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A. et al. (2018) Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol Cell, 71, 858-871 e858.

[196] Slany, R.K. (2009) The molecular biology of mixed lineage leukemia. Haematologica, 94, 984.

[197] Rotem, A., Ram, O., Shoresh, N., Sperling, R.A., Goren, A., Weitz, D.A. and Bernstein, B.E. (2015) Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nature biotechnology, 33, 1165-1172.

[198] UniProt C. UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res 2019;47:D506-D515.

[199] Go N. Theoretical studies of protein folding, Annu Rev Biophys Bioeng 1983;12:183-210.

[200] Alm E, Baker D. Matching theory and experiment in protein folding, Curr Opin Struct Biol 1999;9:189-196.

[201] Shoichet BK, Baase WA, Kuroki R et al. A relationship between protein stability and protein function, Proceedings of the National Academy of Sciences 1995;92:452-456.

[202] Rives A, Meier J, Sercu T et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc Natl Acad Sci U S A 2021;118.

[203] Hu G, Yan W, Zhou J et al. Residue interaction network analysis of Dronpa and a DNA clamp, Journal of theoretical biology 2014;348:55-64.

[204] Heal JW, Bartlett GJ, Wood CW et al. Applying graph theory to protein structures: an Atlas of coiled coils, Bioinformatics 2018;34:3316-3323.

[205] Zhou J, Yan W, Hu G et al. Amino acid network for the discrimination of native protein structures from decoys, Current Protein and Peptide Science 2014;15:522-528.

[206] Vendruscolo M, Dokholyan NV, Paci E et al. Small-world view of the amino acids that play a key role in protein folding, Physical Review E 2002;65:061910.

[207] Del Sol A, O'Meara P. Small-world network approach to identify key residues in protein–protein interaction, Proteins: Structure, Function, and Bioinformatics 2005;58:672-682.

[208] Dokholyan NV, Li L, Ding F et al. Topological determinants of protein folding, Proceedings of the National Academy of Sciences 2002;99:8637-8641.

[209] Jung J, Lee J, Moon HT. Topological determinants of protein unfolding rates, Proteins: Structure, Function, and Bioinformatics 2005;58:389-395.

[210] Bagler G, Sinha S. Assortative mixing in protein contact networks and protein folding kinetics, Bioinformatics 2007;23:1760-1767.

[211] Cusack MP, Thibert B, Bredesen DE et al. Efficient identification of critical residues based only on protein structure by network analysis, PLoS One 2007;2:e421.

[212] Gligorijević V, Renfrew PD, Kosciolek T et al. Structure-based protein function prediction using graph convolutional networks, Nature communications 2021;12:1-14.

[213] Li Y, Wen Z, Xiao J et al. Predicting disease-associated substitution of a single amino acid by analyzing residue interactions, BMC bioinformatics 2011;12:1-9.

[214] Park K, Kim D. Structure-based rebuilding of coevolutionary information reveals functional modules in rhodopsin structure, Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 2012;1824:1484-1489.

[215] Aibar, S. et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nature Methods, 14(11), 1083–1086.

[216] Matsumoto, H. et al. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics (Oxford, England), 33(15), 2314–2321.

[217] Berman HM, Westbrook J, Feng Z et al. The protein data bank, Nucleic acids research 2000;28:235-242.

[218] Hammond DK, Vandergheynst P, Gribonval R. Wavelets on graphs via spectral graph theory, Applied and Computational Harmonic Analysis 2011;30:129-150.

[219] Ganesan K, Kulandaisamy A, Binny Priya S et al. HuVarBase: A human variant database with comprehensive information at gene and protein levels, PLoS One 2019;14:e0210475.

[220] Yu L, Zhang Y, Gutman I et al. Protein sequence comparison based on physicochemical properties and the position-feature energy matrix, Scientific reports 2017;7:1-9.

[221] Kawashima S, Pokarowski P, Pokarowska M et al. AAindex: amino acid index database, progress report 2008, Nucleic acids research 2007;36:D202-D205.

[222] Capra JA, Singh M. Predicting functionally important residues from sequence conservation, Bioinformatics 2007;23:1875-1882.

[223] Roberts K, Alberts B, Johnson A et al. Molecular biology of the cell, New York: Garland Science 2002;32.

[224] Kramer RM, Shende VR, Motl N et al. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility, Biophysical journal 2012;102:1907-1915.

[225] Hou Q, Bourgeas R, Pucci F et al. Computational analysis of the amino acid interactions that promote or decrease protein solubility, Scientific reports 2018;8:1-13.

[226] Lawson DD, Ingham J. Estimation of solubility parameters from refractive index data, Nature 1969;223:614-615.

[227] Bhandari BK, Gardner PP, Lim CS. Solubility-Weighted Index: fast and accurate prediction of protein solubility, Bioinformatics 2020;36:4691-4698.

[228] M Ashraf G, H Greig N, A Khan T et al. Protein misfolding and aggregation in Alzheimer's disease and type 2 diabetes mellitus, CNS Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS Neurological Disorders) 2014;13:1280-1293.

[229] Gromiha MM, Thangakani AM, Selvaraj S. FOLD-RATE: prediction of protein folding rates from amino acid sequence, Nucleic acids research 2006;34:W70-W74.

[230] Chang CCH, Tey BT, Song J et al. Towards more accurate prediction of protein folding rates: a review of the existing web-based bioinformatics approaches, Briefings in bioinformatics 2015;16:314-324.

[231] Srivastava D, Bagler G, Kumar V. Graph Signal Processing on protein residue networks helps in studying its biophysical properties, bioRxiv 2021.

[232] Tixier AJ-P, Nikolentzos G, Meladianos P et al. Graph classification with 2d convolutional neural networks. In: International Conference on Artificial Neural Networks. 2019, p. 578-593. Springer.

[233] Corbo V, Ritelli R, Barbi S et al. Mutational profiling of kinases in human tumours of pancreatic origin identifies candidate cancer genes in ductal and ampulla of vater carcinomas, PLoS One 2010;5:e12653.

[234] Lisabeth EM, Fernandez C, Pasquale EB. Cancer somatic mutations disrupt functions of the EphA3 receptor tyrosine kinase through multiple mechanisms, Biochemistry 2012;51:1464-1475.

[235] Iqbal, M.; Verrall, R. E. Implications of protein folding. Additivity schemes for volumes and compressibilities. J. Biol. Chem. 1988, 263, 4159-4165.

[236] Zimmerman, J. M., Eliezer, N., Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. Journal of Theoretical Biology, 21(2), 170–201. doi:10.1016/0022-5193(68)90069-6

[237] WAUGH, D. F. (1959). Rev. mod. Phys. 31, 84.

[238] SORM, F. (1961). CON. Czechoslov. Chem. Commun. 26, 303.

[239] "Protein, Amino Acid, and Peptides", Reinhold, New York (1943)

[240] Vihinen, M., Torkkila, E., Riikonen, P. (1994). Accuracy of protein flexibility predictions. Proteins: Structure, Function, and Genetics, 19(2), 141–149.

[241] Kyte, J., Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology, 157(1), 105–132.

[242] McMeekin, Thomas L., Merton L. Groves, and Norbert J. Hipp. "Refractive indices of amino acids, proteins, and related substances." 1964. 54-66.

[243] Gromiha, M. Michael, and David AD Parry. "Characteristic features of amino acid residues in coiled-coil protein structures." Biophysical chemistry 111.2 (2004): 95-103.

[244] Venkatachalam CM. (1968). "Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units". Biopolymers. 6 (10): 1425–36.

[245] Capra JA and Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics, 23(15):1875-82, 2007.

[246] Shi, Jianbo, and Jitendra Malik. "Normalized cuts and image segmentation." IEEE Transactions on pattern analysis and machine intelligence 22.8 (2000): 888-905.

[247] Mishra, Shreya, Divyanshu Srivastava, and Vibhor Kumar. "Improving gene network inference with graph wavelets and making insights about ageing-associated regulatory changes in lungs." Briefings in bioinformatics: bbaa360.

[248] Mishra, Shreya, Neetesh Pandey, Smriti Chawla, Debarka SenGupta, Kedar Nath Natrajan and Vibhor Kumar. "A search-engine for single-cell epigenome profiles for multi-purpose applications." bioRxiv (2021). (Under Review)

[249] Mishra, Shreya, Neetesh Pandey, Atul Rawat, Divyanshu Srivastava, Arjun Ray and Vibhor Kumar . “An explainable model using Graph-Wavelet for predicting biophysical properties of proteins and measuring mutational effects” (Under review)

LIST OF PAPERS BASED ON THESIS

1. Mishra, Shreya, Divyanshu Srivastava, and Vibhor Kumar. "Improving gene network inference with graph wavelets and making insights about ageing-associated regulatory changes in lungs." Briefings in bioinformatics: bbaa360.

2. Mishra, Shreya, Neetesh Pandey, Smriti Chawla, Madhu Sharma, Omkar Chandra, Indra Prakash Jha, Debarka SenGupta, Kedar Nath Natarajan, and Vibhor Kumar. "Matching queried single-cell open-chromatin profiles to large pools of single-cell transcriptomes and epigenomes for reference supported analysis." Genome Research 33, no. 2 (2023): 218-231.

3. Mishra, Shreya, Neetesh Pandey, Atul Rawat, Divyanshu Srivastava, Arjun Ray and Vibhor Kumar . "An explainable model using Graph-Wavelet for predicting biophysical properties of proteins and measuring mutational effects" (Under review)

Additional Publications

1. Sharma, M., Jha, I. P., Chawla, S., Pandey, N., Chandra, O., Mishra, S., Kumar, V. (2022). Associating pathways with diseases using single-cell expression profiles and making inferences about potential drugs. Briefings in Bioinformatics.

2. Chandra, Omkar, Madhu Sharma, Neetesh Pandey, Indra Prakash Jha, Shreya Mishra, Say Li Kong, and Vibhor Kumar. "Inferring functions of coding and non-coding genes using epigenomic patterns and deciphering the effect of combinatorics of transcription factors binding at promoters." bioRxiv (2022).

3. Sharma, Rachesh, Neetesh Pandey, Aanchal Mongia, Shreya Mishra, Angshul Majumdar, and Vibhor Kumar. "FITs: Forest of imputation trees for recovering true signals in single-cell open chromatin profiles." NAR Genomics and Bioinformatics 2, no. 4 (2020): lqaa091.

4. Pandey, Neetesh, Omkar Chandra, Shreya Mishra, and Vibhor Kumar. "Improving chromatin-interaction prediction using single-cell open-chromatin profiles and making insight into the cis-regulatory landscape of the human brain." Frontiers in Genetics 12 (2021): 738194.

5. Jha, Indra Prakash, Shreya Mishra, Neetesh Pandey, and Vibhor Kumar. "Stratified assessment for geriatric mental health using probabilistic graphical model: a cross-sectional observational study in India." medRxiv (2023): 2023-03.

6. Pandey, Neetesh, Madhu Sharma, Arpit Mathur, George Anene Nzelu, Muhammad Hakimullah, Indra Prakash Jha, Omkar Chandra, Shreya Mishra, Ankur Sharma, Roger Foo, Amit Mandoli, Ramanuj DasGupta, Vibhor Kumar. "Deciphering the phenotypic heterogeneity and drug response in cancer cells using genome-wide activity and interaction of chromatin domains." bioRxiv (2023): 2023-01.