Approaches for Mapping Unique Phenotype Screens To a Genetic Interaction Network

in partial fulfillment of the requirements for the degree of

Master of Science
in Computer Science

at the Delft University of Technology
to be defended publicly on September 8 2023

Supervisor:	Prof. Dr. Lodewyk F.A. Wessels	TU Delft, Netherlands Cancer Institute
Daily Supervisor:	Dr. Guizela H. Prince	Netherlands Cancer Institute
Thesis Committee:	Dr. Joana de Pinho Gonçalves	TU Delft
Thesis Committee:	Dr. Zhengjun Yue	TU Delft

by

Bram Pronk
4613066

TU Delft

NETHERLANDS
CANCER
INSTITUTE
ANTONI VAN LEEUWENHOEK

Preface

This thesis document signifies the culmination of my efforts towards obtaining the degree of Master of Science in Computer Science, in the Artificial Intelligence track and Bioinformatics specialization, at the Delft University of Technology in The Netherlands. The aim of this research is to extensively explore the compendium of haploid phenotype screens, in order to see its merits toward uncovering genome-wide phenotypic associations in humans. Using computational methods to create networks of interactions between genes, we hope to map interactions that are hard to uncover using more conventional screening techniques.

I was fortunate enough to perform this work under supervision of the Computational Biology group led by Lodewyk Wessels at the Netherlands Cancer Institute (NKI). The exposure to diverse perspectives and cutting-edge research methodologies offered through the institute has equipped me with a much broader perspective and set of tools to tackle challenges in the bioinformatics field. This is in no small part due to the privilege to attend weekly lab meetings and paper discussions, I'd like to thank all the members from the Computational Biology group for their willingness to teach and include me as a member of the group. Furthermore, I would like to extend my gratitude to the academics at the Brummelkamp group at the NKI. Thank you for your willingness to reason with me about network behaviour and making keen suggestions on avenues to explore during my internship. I value my time in group meetings where I could learn from the way you conduct science and the literature discussions have broadened my perspective on genetics research. From TU Delft I'd like to thank Joana Gonçalves and Zhengjun Yue as esteemed members of my thesis committee, and look forward to our discussions.

I am deeply thankful for the support I have received during my time at the NKI. The thesis period was also very much about learning to deal with challenging personal problems and learning to find peace with difficult situations. I could not have done so while working on this project without the support of Lodewyk, Guizela and Bram. Their unwavering support made me feel like I could be honest about the process and always motivated me. They have shown me how to conduct terrific science and how to kindheartedly engage when struggling. I am grateful and hope to take these lessons with me as I aim to continue my path in science. Finally to my friends at the university, thanks for your kindness and support over these past years, I have always felt appreciated and inspired by you.

Bram Pronk
Amsterdam, August 2023

ABSTRACT

Targeted and successful cellular therapies for disease treatment require an extensive mapping of the complex structure and dynamics of molecular mechanisms which determine the behaviour and function of cell. CELL-seq is a genome-wide screening procedure measuring specific and targeted protein quantities as phenotypic readouts and is employed by the Netherlands Cancer Institute to analyze which genes regulate the protein state of interest. This research aims to explore the current compendium of CELL-seq screens that investigate a range of phenotypes, to create a mapping of gene-gene associations that share similar phenotypic profiles and elucidate biology that is hard to uncover with more conventional screening techniques.

We perform exploratory research to investigate the ability of the screen compendium to show network structures that reflect known biological processes. We find that with stringent requirements on interactions the screen compendium shows enrichment for a wide range of biological processes and known protein-protein interactions. We further conclude that the experimental design biases network behaviour and needs to be accounted for when constructing networks. We recommended a mutual k-nearest neighbor network construction approach, which yielded networks with the most biological relevance. We compare the CELL-seq screens using findings from the approaches to the DepMap dataset, a well-known collection of synthetic lethality CRISPR screens, and find that the behaviour of these datasets is in many ways mirrored. We conclude that this is both due to the biology they represent and the differences in the number of screens in each dataset. Finally, we compare the coverage of biological processes between the HAP1 compendium and DepMap, and show large overlap in their coverage. Nonetheless, the differences they do show leads us to bring forward two hypotheses for gene-gene interactions that score strongly uniquely in the CELL-seq networks which are biologically plausible but are not found in DepMap or curated literature, warranting future investigations.

All code pertaining to the methods and figures in this work are hosted on GitLab by the High Performance Computing Facility of the Netherlands Cancer Institute. As such the code can be viewed by supervisors, but further details could be shared upon request.

CONTENTS

1

Introduction

In the post-genomic era, the growing comprehension of cellular biology is catalyzing a transformative impact on medicine. Conditions ranging from Alzheimer's disease and viral infections to aging and cancer are undergoing reevaluation as outcomes of aberrant cellular behaviors, remodelling them as promising targets for cellular therapies including cellular manipulation [1]. Successful treatment necessitates a mapping of the complex structure and dynamics of molecular mechanisms and processes which determine the behaviour and function of cells [2], [3]. To this end, associating DNA sequences (genotype) to phenotypic outcomes elucidates this molecular architecture and remains one of the key questions in genetics and systems biology [4], [5].

It is well-established that while most phenotypes exhibit strong genetic associations, they often explain only a fraction of the total genotypic variance [6]. Instead, a wide variety of phenotypes are subject to polygenic effects from multiple genes spread widely across the genome [7]. As such, the identification of each gene's role in cellular pathways and its associations with phenotypes heavily depends on understanding its contextual co-functioning with other genes [8]. These functional relationships can be modelled as a network of molecular interactions, where genes are represented by nodes and edges between genes signify a relationship or interaction. Such a network is referred to as a gene-gene interaction network. The nature and strength of these interactions is systematically established by the data-collection technique employed. A variety of high-throughput measurement techniques have taken advantage of these powerful network models; pairwise gene knockout screens of yeast have been used to create a map of genetic interactions for the model organism [9], protein-protein interaction (PPI) databases to retrieve the PPI network associated with hereditary disorders [10] and genome-wide CRISPR/Cas9 screens for mapping genetic interactions in cancer cells [11].

While these approaches show promising results, each suffers from limitations that propagate from the data collection technique into the produced gene-gene interaction network. Reductionist biology using pairwise gene knockout screens has provided a foundation of genetic interactions, but fails to capture the polygenic origin of traits and scales poorly to the circa 20.000 protein-coding genes in the human genome. Genome-wide CRISPR/Cas9 screens often use high-level phenotypes such as growth or cell fitness, which is problematic as not all cellular events affect these phenotypes [12]. In addition, genes acting in completely different mechanisms can be wrongly associated by the high-level phenotype, which requires follow-up studies to deconvolve these associations [13]. And gene expression measurements by RNA-seq suffer from the fact that expression needs to be averaged across cells which leads to the inadvertent capturing of cellular heterogeneity in the data [14] and it does not capture cellular events which do not have transcriptional effects [12].

1.1. Haploid Phenotype Screens

CELL-seq is a measurement technique currently used by the Brummelkamp group at the Netherlands Cancer Institute which aims to overcome the aforementioned limitations of traditional assay techniques. The central conviction is in concordance with genetic perturbation approaches: inducing mutations and measuring their influence on biological processes is the most powerful unbiased approach linking genotype to phenotype [15]. However, generating and recovering bi-allelic mutants in (human) diploid cells poses a challenge, as the inactivation of a single allele can remain masked [16]. Furthermore, the presence of conflicting instructions within two alleles of the same gene adds another layer of complexity to the genotype-phenotype relationship [17].

CELL-seq overcomes this difficulty by employing a near-haploid cell line termed HAP1 [18]. Haploidy means there is only a single copy of the genome present in the cell. While haploidy in human cells is normally limited to gametes, the HAP1 cell line is derived from a chronic myeloid leukemia patient after reprogramming experiments to create induced pluripotent stem cells. The karyotype for HAP1 cells is shown in Figure 1.1. The haploid nature of the cell line makes the insertion of a termination sequence into the genome, gene-trapping, result in a mutation of the gene that prematurely truncates its RNA transcript making the gene considered nonfunctional. This process is called gene-trap mutagenesis. In a CELL-seq screen, a large population of cells is subjected to random gene-trap mutagenesis such that one random gene is inactivated per cell, after which a quantitative measure of a protein state is measured in each cell. The exact details of the screening procedure and implications are further discussed in the Methodology, but this method allows for the direct linkage of genetic mutations for every gene inactivated in the population to a protein state. A protein state serves as a low-level phenotypic readout, meaning much more specificity to a cellular mechanism, whereas traditional screening methods with high-level phenotypes measure much broader categorized phenomena like cellular growth.

167 CELL-seq screens have been accumulated by the Brummelkamp group to create a haploid phenotype screen compendium which covers a wide range of diverse low-level phenotypes. The screens accomodate for the polygenic origin of traits, scale to the entirety of the protein-coding genes in the human genome, use a wide range of low-level phenotypes to capture very specific phenotypic regulators and are reduced in noise through its use of a single cell line in all screens. Decades of work in both yeast as a model organism and in human cancer cells shows that genes with similar phenotypic associations for high-level phenotypes such as growth or synthetic lethality can be identified as functionally similar genes through guilt by association [19], [20]. In this work,

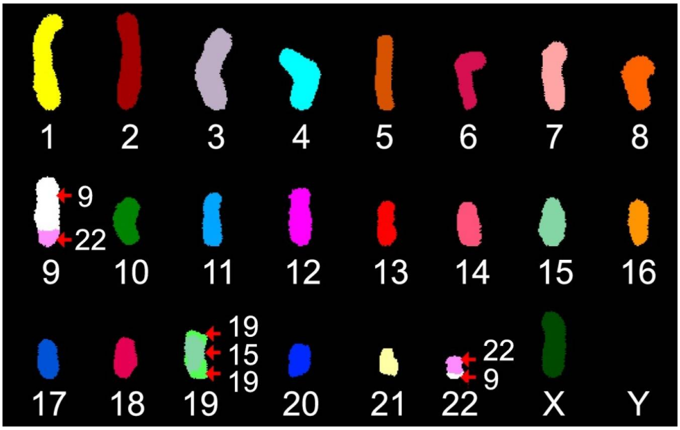


Figure 1.1: Spectral karyogram of the HAP1 cell line, showing its near-haploid composition. Notably, Chromosome 15 retains two copies, one of which is fused to Chromosome 19. Also note the reciprocal translocation of genetic material between Chromosome 9 and Chromosome 22, a hallmark of the original KBM-7 cell line. Source: Figure adapted from Figure 1a in [21].

we explore CELL-seq as a technique to measure gene-gene interactions, and similarly to established work on traditional genome-wide assays apply computational techniques to create a gene-gene interaction network.

1.2. RESEARCH OBJECTIVES

The aim of this research is to extensively explore the compendium of haploid phenotype screens, in order to see its merits toward uncovering genome-wide phenotypic associations in humans. Using computational methods to create networks of interactions between genes, we hope to map interactions that are hard to uncover using more conventional screening techniques. This will be the first work integrating data of this compendium for the purpose of creating a gene-gene interaction network. As such the work is exploratory in nature, to establish whether known techniques are directly applicable to the dataset and systematically exploring interpretable and novel approaches suited to the characteristics of the dataset. The individual screens were performed to test hypotheses related to the measured phenotype, and not originally designed for their integration leading to network construction. With the integration of all screens, we bring forward a methodology on how to engineer a solution for a frequent problem in bioinformatics, the disparity between original experimental design and required data for interaction modelling [22].

Towards this goal, we formulated the following research questions.

1. *Can we leverage the dataset to create a gene-gene interaction network that shows proof of relevancy in current literature on genetic interactions?*
   This study defines gene interactions through phenotypic associations from the unique screening procedure of CELL-seq. From this we aim to construct a network of genes that displays the molecular relationships on the basis of a varied range of phenotypes. We then ask if connectivity patterns in the network relates clearly to biological phenomena [23]. Using literature on known molecular interactions of biological systems, protein-protein interactions (PPIs) and protein complex co-membership we show how constructed networks recapitulate real biology. From this map, we can extract interactions not established in literature, but apparent from the screens and network methodology, in order to bring forward hypotheses for new biology.

2. *How does the haploid phenotype screen compendium compare to DepMap, whose dataset is antagonistic to ours by screening one phenotype over more than a thousand cell lines?*
   The Cancer Dependency Map Project (DepMap) [24] is a large study and dataset on dependency profiles of genes in currently over one thousand cancer cell lines from 31 types of cancer. The dataset consists of a collection of CRISPR screen that precisely and systematically inactivate genes across the genome and measure the effect of each perturbation on cell fitness. The large and growing dataset has led to its inclusion in many reports that show shared functionality of genes based on associated fitness profiles across cell lines. A very interesting juxtaposition to the HAP1 screens, is that in DepMap the same phenotype (fitness) is covered over many cell lines, while the HAP1 screens cover many phenotypes (protein state readouts) but in a single cell line. Contrasting and comparing to this dataset can provide meaningful insights into the relevancy of the HAP1 dataset.

3. *Does the limited compendium screen size fit the data-hungry task of cellular mapping?*
   The HAP1 screen compendium has been added to over several years and currently contains 167 screens for a varied selection of phenotypes. It is however important to acknowledge that there might be an insufficient number of screens to accomplish the ambitions of creating a comprehensive biological network. A low number of screens reduces the statistical power, and therefore hampers accuracy of meaningful associations between genes and could lead to false negatives.

   (a) *How can we optimally add screens to most effectively increase diversity in the phenotypic space?* Being limited in the number of screens also leads to a more narrow scope of analysis: currently screened phenotypes may not capture the full diversity of biological contexts, leading to a limited representation of gene-gene interactions and functional associations. Consequently, our analyses might overlook crucial interactions relevant in specific cellular contexts. Future work on the screen compendium aims to expand its library of screens to combat this. By investigating the current coverage of biological contexts, we aim to propose phenotypes for future screens that most efficiently expand coverage while locking in on most relevant research avenues brought forward by the results of RQ1.

1.3. OUTLINE

This thesis work is structured as follows. First, the Methodology covers the pipeline used in this study to answer the research questions outlined above. The visual representation of the pipeline is depicted in Figure 1.2, illustrating the several stages of this investigation which the Methodology will cover in order. First, a more in-depth analysis of the HAP1 dataset discussing general features introduced through the screening method. The HAP1 screen compendium's creation is established in literature, but as this work introduces it for this particular usage, it is important to be extensive in its description, advantages and limitations. We further elaborate on the secondary contrasting dataset used in this work for comparative analysis: DepMap. Next we cover how the datasets are pre-processed in order to accommodate for two of the major biases in the HAP1 screen compendium, which lead to the pairwise distance matrix for genes used in this research. Following this, we introduce networks, how they are validated and scored for biological relevance using literature and external databases, and how their characteristics are defined and measured. To end the Methodology, background information is provided on used methods to answer the research questions that are presented in the Results Section 3. The Results section refers to this background material when more detail is appropriate. The findings in the Results are introduced mostly in chronological order of the project's completion. The Conclusion summarizes this works main contributions by highlighting and answering the research questions. After this summarizing part, a critical discussion on this work's limitations is provided and finally recommendations are made for future work.

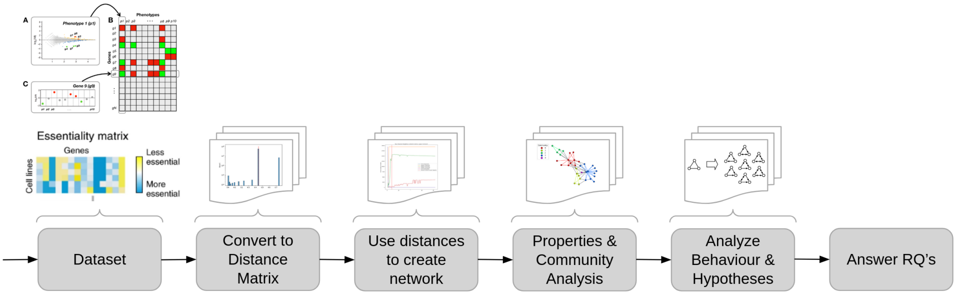


Figure 1.2: Overview of the network construction and analysis pipeline of the project.

2

Materials and Methods

2.1. Datasets

2.1.1. HAP1 Phenotype Screen Compendium

The exact CELL-seq procedure is described by Brockmann et al. in [12]. We include an outline of the procedure in the following paragraph since it is important to understand the advantages, limitations and biases discussed later in this work. The outline is visualized in Figure 2.1 adapted from Brockmann et al.

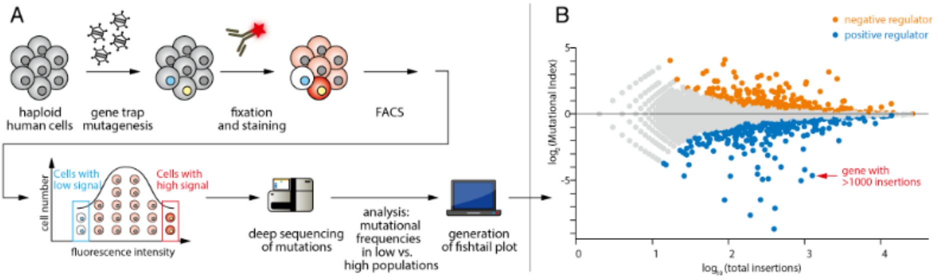


Figure 2.1: (A) The pipeline shows a large population of HAP1 cells used to obtain mutational index values for a specific phenotype. The values and associated statistical significance are shown in the fish-tail plot in (B).

A large population of 10^8 HAP1 cells is subjected to randomized gene-trap mutagenesis, by inserting one termination sequence per cell randomly in the genome. The termination sequence signals the transcriptional complex to release the RNA transcript, now prematurely, voiding the gene product and thereby nullifying gene expression. This induces a loss-of-function mutation on a single gene per cell. This population of inserted cells is expanded, then fixed to preserve and to halt ongoing chemical reactions, and finally permeabilized to allow accessing intracellular molecules. These steps allow an antibody targeting a specific protein of interest, which will be the phenotypic readout. The antibody contains a fluorescent stain, so the measurement of protein state abundance can be performed using flourescence-activated cell sorting (FACS) [25]. This fluorescense-level sorted population of cells can be split into categories, such that the least flourescent cells and most fluorescent cells are categorized as a low- and high population, meaning they have either a very low or very high abundance of the targeted protein. In the context of the HAP1 dataset, we use the terms high and low *population* and high and low *channel* synonymously. After performing a deep sequencing of mutations on both the low and high populations, the frequency of mutated genes can be compared between them. This ratio of mutations in both populations is expressed by the Mutational Index (MI). We use the same definition for MI as Brockmann et al. and include the equation from [12] in Equation 2.1 here for completeness.

$$MI = \frac{G_H/O_H}{G_L/O_L} \tag{2.1}$$

Where G_H is the number of insertions within a specific gene in the high channel, O_H is the total number of insertions that are not in that specific gene in the high channel, and G_L and O_L are similarly defined but for the low channel. In other words, it is the ratio between insertions in a gene in the low and high population, normalized by the ratio between the total amount of out-of-gene insertions in the low and high populations.

To determine whether genes are significantly enriched for disruptive gene-trap integrations in either low or high channel, the number of insertions in that gene in both channels were counted as well as the number of total out of gene insertions in both of the two populations. These four values serve as inputs to a 2x2 contingency table (Table 2.1). A two-sided Fisher's exact test was used to determine if there was a significant association between the low and high channels for every gene in every screen. The result of the test is the odds ratio, indicating the strength and direction of the association, and the corresponding P value. Note that the odds ratio is equal to the definition of MI in Equation 2.1. The final MI values used in this work are scaled with $\log_2(MI)$. The resulting P values are then adjusted for multiple testing with the Benjamini-Hochberg false discovery rate correction. In summation, a high MI (> 1) determines a negative regulator for a screen phenotype readout, since this means there are more insertions into the gene in the high population (disabling expression) and fewer in the low channel comparatively. The converse (MI < 1) is true for positive regulators. Figure 2.1B shows the result of an example screen, with the $\log_2(MI)$ values on the y-axis and the total insertions ($\log_2$) on the x-axis. Each dot represent one gene of the screen, and genes with a significant P value from the two-sided Fisher's exact test are colored as significant (positive or negative) regulators.

This procedure was performed for 167 screens targeting around 100 different pheno-

	High	Low
Gene insertions	G_H	G_L
Out of gene insertions	O_H	O_L

Table 2.1: Contingency table showing the frequency distributions of in- and out-of gene insertions in both the high and low populations. This contingency table for every gene was analysed for statistical significance between the high and low population using a Fisher's exact test.

types. For the remainder of work, we use the following definition for the HAP1 dataset. The HAP1 phenotypic screens can be described as a data matrix $\mathbf{X_H} = \underset{19385\times167}{\mathbf{X_H}}$ for 19385 genes measured over 167 phenotype screens. Each entry in $\mathbf{X_H}$ contains the $\log_2(MI)$ value. We can similarly define such a matrix holding the corresponding P value from the Fisher's exact test for each gene in each screen as defined above.

Related to the P values, when using a significance threshold of $\alpha < 0.05$, we define the term *score* for a gene indicating that when a gene has scored in a screen, it means its P value was below the significance threshold α. Out of all genes, 11093 genes are determined to score in at least one screen. Out of those 11093 genes, there is a wide variety of significant MI values; some gene profiles are varied across all phenotypes, some only for a specific range and some are highly specific and score for only one phenotype.

In practice, it is generally acknowledged by the biologists conducting and interpreting the phenotype screens that an absolute $\log_2(MI)$ value of 1 is considered high. A $\log_2(MI)$ of $|1|$ means the ratio between high and low channel insertions in a gene is equal to the rate of high and low insertions out of the gene. Out of all genes, 17471 genes have a $\log_2(MI) \geq |1|$ in at least one screen. As such, setting a pre-defined threshold for an MI score alone is not fully informative, as there are genes that pass this threshold without being considered statistically significant (only 11093 are in at least one screen). Therefore it is required to use thresholds for both the effect size as well as the P value.

A study by Turco et al. that collected and compared genome-wide phenotypic screens in *Saccharomyces cerevisiae* uses another metric to determine genes that are strongly associated to a phenotype [4]. They introduce the concept of a normalized phenotypic value (NPV), where each gene's value in the screen is expressed by a modified z-score standardization where instead of the mean the mode is used as a reference. From these scores, they determine a gene to be strongly associated to a phenotype if $|NPV| \geq 3$. Due to the different screening procedures between those used in the study by Turco et al. and the HAP1 screens, and adding the different effect sizes of each screen in the HAP1 dataset, the NPV is not directly applicable here since almost all genes have only non-recurring MI values and thus the mode almost always occurs only once. Instead of the mode, a more applicable metric would be to use standard z-scoring with the mean as reference, which leads to 10503 genes having a z-score $\geq |3|$ in at least one screen, with only 4955 of them also being considered statistically significant through the corresponding P value.

2.1.2. DepMap: Synthetic Lethality genome-wide CRISPR screens

The Achilles project is a collaborative initiative aimed at unraveling the genetic determinants of cancer vulnerabilities. It encompasses a vast collection of molecular and pharmacological data from numerous cancer cell lines, representing a diverse range of cancer types and genetic backgrounds. By perturbing genes and observing their impact on cell survival and growth, the Achilles project identifies genes essential for cancer cell viability, which is offered as a valuable resource for the identification of potential therapeutic targets in a cancer dependency map dubbed DepMap [24]. This work uses the 22Q4 version. This version is a dataset of genome-scale CRISPR screens from 1078 cell lines containing gene-level essentiality scores for 17456 genes. These essentiality scores are expressed as CERES scores, which are the scores following application of the CERES algorithm on raw measurements of cell proliferation in a genome-scale CRISPR loss-of-function screen, which accounts for some specific biological contexts that lead to more false positives [26]. The dataset was downloaded from `https://depmap.org/portal/download/all/`. The required files are `Model.csv` and `CRISPRGeneEffect.csv`. In a similar fashion to $\mathbf{X_H}$, we use the following definition for the DepMap dataset. The DepMap CRISPR synthetic lethality screens can be described as a data matrix $\mathbf{X_D} = \underset{17456\times167}{\mathbf{X_H}}$ for 17456 genes measured over 1078 CRISPR screens, such that every entry in $\mathbf{X_D}$ contains a CERES score for one particular gene in a screen. DepMap uses OncoTree [27], a widely-adopted tumor classification system to identify tumor types with cell lines. The 22Q4 version captures 31 distinct lineages, and 202 cancer subtypes. Most represented cancer types are lung (236), lymphoid(208) and CNS/brain (125).

We opted for DepMap for constructing gene-gene interaction networks for four reasons. First and foremost, it is in a sense antagonistic to the HAP1 screen compendium, as DepMap covers one phenotype but over a variety of cell lines, where HAP1 covers multiple phenotypes over one cell line. This makes comparing and contrasting interesting, as we can now compare results to a more traditional approach to see if $\mathbf{X_H}$ really tackles limitations of high-level phenotypes. Secondly, the dataset is meticulously curated, containing robust and reliable gene dependency scores derived from rigorous experimental techniques. The project was continuously improved over six years to accomplish this. Thirdly, DepMap includes a broad spectrum of cancer cell lines, representing various tissue origins and genetic alterations. This coverage of a wide range of biology makes findings across cell lines more robust and conserved, while also allowing research into specific cancer or tissue types. Finally, the dataset is publicly accessible, and has been used for several successful realizations of clustering and network approaches [28], [29]. The research by Weinberg et al. identifies a problem with the DepMap data [28] which translates very well to the HAP1 dataset, the application of their proposed solution on the HAP1 screen compendium and on the DepMap dataset is described in the following section. The application of this technique to both datasets makes comparisons between the two fairer.

2.2. Dataset Preprocessing

2.2.1. Generalized Least Squares

It is important to note the context in which these screens were performed, as the screens in $\mathbf{X_H}$ were not done with this research plan in mind, where a network of interactions is derived. Instead, they originate from various research efforts with different end-goals. The screens cover various quantitative cellular traits in multiple areas of biology, but they do so non-uniformly. Some examples include lipid transport, which has been mapped by 9 screens; phospho-S6 ribosomal protein as a marker for cell signalling pathways that respond to both growth factors and the nutritional state of the cell [30] has 10 screens; and p38 mitogen-activated protein kinases are covered by 3 screens. 34 phenotypes are covered by only one screen. One of the implications of this varied spread of screen-to-phenotype coverage is that it introduces strongly related readings, and therefore the features show nonindependence. For example, genes that are strong regulators of the p-S6 phenotype are going to have inflated correlations since their activity was explicitly measured so frequently.

We have previously described the DepMap data used for comparing and contrasting the HAP1 dataset. Similarly to HAP1, the DepMap data also shows nonindependence. In DepMap, this arises due to the origin of certain cell lines from shared tissues or lineages. In 2021, Wainberg et al. tackled the issue of nonindependence in DepMap by employing a methodology based on generalized least squares (GLS) [28]. This approach explicitly considers the nonindependence of cell lines, ensuring accurate analysis and interpretation of the data. Introduced by Aitken in 1935 [31], Generalized Least Squares (GLS) constitutes a technique for parameter estimation within a linear regression model. Its application is designed for application when the residuals of a regression model exhibit a discernible level of covariance or correlation, which violates a core assumption of ordinary least squares. The implementation used by Weinberg et al. incorporates a computationally efficient method to solve GLS, described in the methodology of the work by Weinberg et al. As such, we used the publicly available code [1] of [28] verbatim to ensure reproducibility.

We ran GLS on each gene pair in $\mathbf{X_H}$ and separately also for $\mathbf{X_D}$, thus resulting in a distance matrix $\mathbf{D} = \underset{X_n \times X_n}{\mathbf{D}}$ of GLS P values where X_n represents the number of genes in each dataset $\mathbf{X}$. We use the notation $\mathbf{D_H}$ for the distance matrix of the HAP1 dataset and similarly use $\mathbf{D_D}$ for the DepMap dataset. The GLS P values represent the statistical significance of the null hypothesis where the coefficient is equal to 0. Interactions lower than the predetermined $\alpha < 0.05$ are significant. In concordance with the work by Weinberg et al., we applied Benjamini-Hochberg FDR correction on the GLS P values. For every gene-gene pair, GLS also provides the sign of the interaction.

Finally, $\mathbf{D}$ was transformed using $-\log_{10}(\mathbf{D})$ such that higher values in $\mathbf{D}$ correspond to relations with higher confidence scores. Due to the size of $\mathbf{D}$, all values were converted from `float64` to `float32`, which due to floating point errors led to some small P values being set to 0. Since $-\log_{10}(0)$ is undefined, those values were transformed to the lowest P value > 0 for that gene. The GLS procedure is visualized in Figure 2.2.

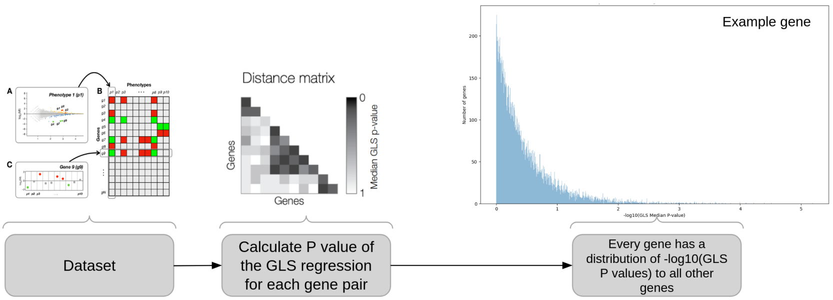

A
Phenotype 1 (p1)
B
Phenotypes
p1 p2 p3 ... p8 p9 p10
g1 g2 g3 g4 g5 g6 g7 g8 g9 ⋮ gN
Genes
C
Gene 9 (g9)
p1 p2 p3 ... p10
Distance matrix
Genes
Genes
0
1
Median GLS p-value
Example gene
Number of genes
0
50
100
150
200
0 1 2 3 4 5
-log10(GLS Median P-value)
Dataset
Calculate P value of the GLS regression for each gene pair
Every gene has a distribution of -log10(GLS P values) to all other genes

Figure 2.2: From dataset $\mathbf{X_H}$ / $\mathbf{X_D}$ to a distance matrix of GLS P values between every gene-gene pair.

[1] https://github.com/kundajelab/coessentiality/tree/master

2.2.2. Monte Carlo Sampling Against Biased Correlations

As discussed, the individual phenotype screens are the result of a large population of cells being subjected to randomized gene-trap mutagenesis. Due this nature of the screens, it is not possible for each screen to include the exact same number of HAP1 cells, nor for every gene to be covered equally by the gene-trapping. As such, screens have different effect sizes and not all genes are scored in every screen. After being sorted into a high and low population, genes that were not inserted into in the high or low population are given a count of 1 in the channel for that population. We use 1 instead of 0 in order to avoid division by 0 errors when calculating the MI using Equation 2.1.

With uninserted genes or with low values for gene insertions, Equation 2.1 is essentially simplified to $\frac{O_H}{O_L}$, which we define as the *screen offset*. A consequence of this is that in each screen many uninserted genes have an MI value equal to the screen offset. When using the MI profiles of MI values across 167 screens to correlate genes, many gene-pairs are correlating almost perfectly; a biased correlation due to a lack of insertions. Since the number of insertions used for the MI can be so low, most of the scores equal to the gene offset have a large associated P value of up to 1. So the biased correlations are high even though the associated confidence intervals (CIs) of each MI are also high.

A proposed solution to lessen the effect of biased correlations by taking the P values of the MIs into account is to use Monte Carlo sampling to determine a MI value for a gene in each screen based on the size of the CI. The procedure is as follows. We begin with a matrix of 19385 genes over 167 screens where each gene thus has 167 MI values, each with their own associated P value based on the number of insertions. The Fisher's exact test doest not provide the width of the CI directly, but we can calculate it with a standard formula. We use the definition of the formula for CI from the book by Tenny and Hoffman, who also indicate it to be standard practice to use the 95% confidence interval [32]. The formula is given in Equation 2.2.

$$CI = exp(\ln(OR) \pm Z \times SE) \tag{2.2}$$

Where OR is the odds ratio from the Fisher's exact test, Z is the critical value from the standard normal distribution corresponding to the desired confidence level, in our case standardly defined to be 1.96 for 95%, and SE is the standard error of the natural logarithm of odds ratio. SE is given by $SE = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$, where a, b, c and d are the counts from the 2x2 contingency table shown in Table 2.1. The ± can be interpreted as a + when calculating the upper 95% CI and as a − when calculating the lower 95% CI, so the final width of the CI is given by the upper CI minus the lower CI. Now that we have the width of the confidence interval, we sample a new MI value by taking one sample from a normal distribution placed on the CI, defined as $\mathcal{N}(MI, \frac{CI}{2*1.96}))$. The result is a data matrix of 19385 genes over 167 screens but now with a newly sampled MI value. We perform this approach 1000 times resulting in 1000 data matrices.

For each of the 1000 data matrices, we run GLS as described in the previous section. So now, for each gene-gene pair we get 1000 GLS P values, each of the P values based on the MI values found in one of the 1000 sampled data matrices. Each GLS P value was scaled with $-\log_{10}(P)$, such that higher values can be interpreted as more significant. We take the median of those 1000 $-\log_{10}$ scaled GLS P values and denote it as the true GLS P value for that specific gene-gene pair, which we henceforth refer to as the median GLS

P value. This approach of calculating the median GLS P values was validated prior to the commencement of this work by showing that meaningful clusters enriched for terms in the Reactome pathway database [33] after projecting the 19385 x 19385 matrix of median GLS P values using UMAP and clustering the UMAP projection using DBSCAN [34]. We do not apply the MC sampling procedure to the DepMap dataset since the above problem of too little insertions leading to biased correlations does not translate to the original CERES score matrix employed by DepMap. As such, the definition of $\mathbf{D_D}$ remains unaltered since its definition in Section 2.3, but $\mathbf{D_H}$ now contains the median GLS P values obtained from the above sampling procedure.

We can visualize the effect of the Monte Carlo sampling by showing the distribution of median GLS P values for each gene to all other genes, or all values from the 19385 x 19385 median GLS P value distance matrix. The distribution is shown in Figure 2.3. We contrast the distribution of the HAP1 dataset based median GLS P values with the 17456 x 17456 GLS P values matrix of the DepMap dataset.

For DepMap, we report a similar distribution to the one in Figure 1b of the work that introduces the GLS approach and applies it to DepMap by Wainberg et al [28]. As specified by their approach, the median of the distribution should be close to 0.5 for a well-calibrated method. They report a median of 0.48 for DepMap, while Figure 2.3 shows a median of 0.5. This is most likely due to the usage of the most recent 22Q4 version of DepMap which more than doubles the number of cell lines compared to the 18Q3 version of [28].

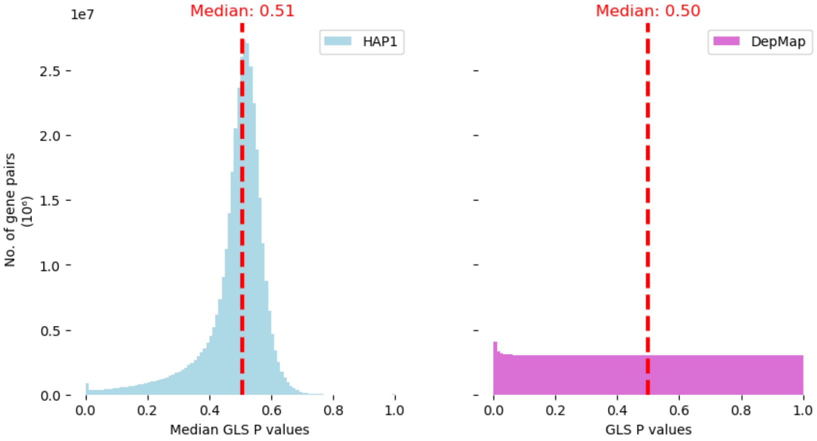


Figure 2.3: Distribution of (median) GLS P values for every gene-gene combination in the HAP1 dataset (left) and DepMap dataset (right). The median is shown for both datasets to show calibration, and does not refer to the median of the MC sampled GLS P values.

The HAP1 dataset distribution does not show an even spread of median GLS P values and more closely resembles a Gaussian. Gene-gene pairs that showed no correlation at first can start to show correlating behaviour by nature of the sampling. Instead of an even spread across the P value range, we see that most gene-gene pairs start to show a P value around 0.5 for their GLS. Due to the nature of the MC sampling, genes with low confidence intervals remain quite similar to the original reported MI values. Correlations of gene-gene pairs where each has a high number of insertions remain fairly stable across samples. On the other end, gene-gene pairs with high correlations due to a lack of insertions leading to biased correlations are much less likely to do so. We show in Chapter 3.4 that no interactions reported in the finalized networks have GLS P values in the range of the distribution's mean, which should reassure that even though the MC sampling signature is present in the distribution, the significant interactions ($P < 0.05$) are those found in the presented networks.

2.3. Network Construction

In this work we use the term network and graph interchangeably. We formally define a graph as $G = (V, E)$, where V is a set of vertices (or nodes) such that every $v \in V$ represents a gene, and E is the set of edges such that every edge $e \in E$ is a set of paired vertices. An edge is weighted according to the GLS P value in the distance matrix $\mathbf{D_H}$ or $\mathbf{D_D}$ depending on the dataset as defined in the Data Preprocessing Section. Biologically an edge represent the weighted degree to which two genes have phenotypic profiles (the MI and P values across screens in $\mathbf{X_H}$ or essentiality scores in $\mathbf{X_D}$) that show a relationship based on the GLS analysis of Section 2.2.1 (and also 2.2.2 for HAP1). Since we lack information about direction of interaction in the distance matrix $\mathbf{D}$, we define G to be an undirected graph, and as such E is the set of unordered paired vertices. As described in previous sections, the HAP1 dataset $\mathbf{X_H}$ was not suited for direct network analysis due to feature nonindependence and biased correlations. Therefore, results presented here are based on the previously defined distance matrix $\mathbf{D}$ unless explicitly stated otherwise.

This work is exploratory in nature, and thus at every point the findings of an analyses informed the next direction of the research. As the story progresses, we present the methods for network construction and the decisions behind each applied method in the Results. However, this section in the Methodology details how networks were analyzed, validated for accuracy and visualized. In some cases it is important to include additional background on the working of algorithms or approaches. Therefore, there is a dedicated 'Background' section that provides additional information further along in the Methods.

2.3.1. Community Finding

We applied community detection to retrieve groups of nodes that exhibit a higher interconnectivity. Nodes in a module work together (interact) to achieve a distinct function [35] or entail a similar process. Genes with similar profiles for broad phenotypes or gene expression measurements often indicate them being active in the same biological processes [36]. Structural analysis can hypothesize which areas of a network form such a community. In this work the terms communities, clusters and modules will refer to the same structural patterns in the context of network analysis, and as such they can be substituted for one another. For community detection we employed the Leiden algorithm [37]. Leiden is closely related to a network property concerning community structures called modularity (Q). Modularity acts as a quantifier, assessing the degree to which nodes within a network form closely-knit clusters that exceed random expectations. A graph with high modularity has a tendency to be grouped into communities, since its subgraphs have more edges than what would be expected by chance. In other words, there are subgraphs that have a significant amount of internal connections and fewer external connections. Modularity is defined to be $-1 \leq Q \leq 1$, with $Q = 1$ indicating a perfect subdivision of non-overlapping communities, Q closer to 0 means subdivision is not significantly better than random, and $Q = -1$ meaning the network is subdivided worse than expected by random chance. The Leiden algorithm treats the community detection problem as a Q maximization problem. As stated by Brandes et al., modularity optimization is a NP-hard problem [38], and therefore the Leiden algorithm is a heuristics algorithm.

In brief, the Leiden algorithm starts with a partition of the network in which each node belongs to a distinct community. Then for each node v_i, the algorithm calculates the change in modularity δQ when moving v_i into the community of another neighbor v_j. v_i is moved to the community that gains the maximum change in δQ. It is therefore a greedy algorithm, randomly starting at a node v_i and making the optimal choice at each iteration. This continues until no further movements improve on Q, after which all communities are bundled into one super-node each which representing each community, called the aggregate network. These super-nodes are connected with edges whose weights are the sum of all edges between the communities before they are represented by super-nodes, after which the algorithm returns to the movement phase until it converges when no moves in the aggregate networks improve Q.

Furthermore, the Leiden algorithm is applicable to weighted networks, and due to it being a modularity optimization algorithm, has very interpretable results. We turned to the Leiden algorithm after its recent recommendation on clustering in single cell approaches [39], and frequent inclusion in other biological networks. The `community_leiden` implementation of the `igraph` package has several hyperparameters. We tested a range of parameters for the resolution (scale of clusters found) and switched between modularity and the Constant Potts Model for the objective function. Final values include a resolution of 1, the modularity as objective functions and running the algorithm until a stable number of partitions was found.

2.3.2. Network Validation

We conducted five assessments of the ability of created networks to recall known biological relationships based on well-established databases for such relationships.

GO:BP Enrichment Analysis

The networks in this work aim to showcase cellular organization through the mapping of phenotypic relationships. Functional annotation of nodes and communities in the network can be used to identify which parts of biology are reflected accurately in the network and how they relate to each other. Additionally, uncharacterized genes that are shown to correspond to the same functional annotation can help elucidate their function, a method extensively used in *Saccharomyces cerevisiae* genetic interactions such as in the SAFE method by Baryshnikova [40]. Functional annotations are provided by the Gene Ontology (GO) database [41], a standard resource to describe the roles of genes in a biological context. The GO database is a vocabulary of associations between genes and GO terms, which are subdivided into three categories biological context: i) molecular function (GO:MF), ii) biological process (GO:BP) and iii) cellular component (GO:CC). In yeast it was shown that genes with similarly phenotypic profiles are frequently co-occurring in the same biological process [42]. Therefore, similar to SAFE and others, this work only uses GO biological process (GO:BP) terms. Besides extensive usage of GO term enrichment analysis in literature, this analysis was also recommended by a 2013 survey on gene regulatory network inference evaluation metrics [43].

The GO:BP terms describe the functional theme associated to a gene and are structured in a directed acyclic graph (DAG) representing a hierarchical ontology structure such that GO terms at the top of the tree describe broad categories of biology to which many genes are associated while terms become more specific towards the leaves. To establish whether a group of genes can be associated to or defined by a GO term, a common approach is to perform gene set enrichment analysis (GSEA). Enrichment is defined as a metric of statistical overrepresentation of a a biological process within a given set of genes when compared to a larger background set of genes. An example would be the set of genes belonging to a retrieved community of a network of the HAP1 screen compendium where the background is all the genes in the network.

Enrichment analysis was performed using the g:Profiler API [44], available at https://biit.cs.ut.ee/gprofiler. g:Profiler was introduced as a tool to standardize enrichment analysis for different databases, and it was used in this work due to its extensive information of evidence of genes belonging to a GO term, and its frequent quarterly updates of databases. To ensure the annotations are of highest quality, we explicitly did not include any electronic annotations of genes to a GO term. Therefore all GO terms found in the network are based on experimental evidence only. One downside to using g:Profiler is that is not clear which genes are used as a background for the enrichment analysis. Therefore the API calls always include the background set of genes (all genes in the network) so that enrichment analysis is accurate, but this does impede on the speed. Because of the time it takes to perform the enrichment, we limited the number of communities found to the 500 largest in each network. From emperical evidence we see that this amount usually captures all the communities in found networks.

g:Profiler calculates the enrichment using a one-sided Fisher's exact test, measuring the significance of the association between the genes in the query set and those belonging to an ontology term. The returned P value is the probability of observing the intersection between the two, plus all possible larger intersections. The P values all undergo multiple testing correction using the g:SCS algorithm, which is the default method for doing such a correction on P values from enrichment analysis. The algortihm was introduced in the work by Reimend et al. along with implementation details [45], but notably it is specifically designed for the ontology structure of GO, and as such is better suited than more traditional over-representation methods like Bonferroni.

STRING Assessment

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) [46] is a widely used and highly reliable resource for exploring protein-protein interactions and functional associations. Besides cluster annotations, STRING provides interaction scores for gene-gene pairs that are determined to have an interaction via evidence originating from computational similarities based on inter-species whole-genome comparison, direct lab assays, presence in other well-known and curated interaction databases and finally co-occurence of proteins in papers published on PubMed (text-mining) [47]. The interactions are scored between the range of 0 and 1000, and are classified by the STRING-specified confidence thresholds (low ≥ 150, medium ≥ 400, high ≥ 700, highest ≥ 900). The approach to GO enrichment is very dependent on the found subsets of genes, found in this work by community detection algorithms. By only considering edges in the network as, STRING validation is thus complementary to GO:BP enrichment analysis as it is cluster-agnostic.

As STRING enrichment in this work is used on edges in the network, this allows for the introduction of measurements on precision and recall of the network. Precision of the network refers to the rate of true positive predictions over all positive predictions: precision $= \frac{TP}{TP+FP}$. The recall, or sensitivity, refers to the rate of true positive predictions over the actual positives: recall $= \frac{TP}{TP+FN}$. In the context of biological networks in this work, a prediction reflects an edge being present between two nodes in the network. For this work, we conduct an analysis of precision-recall curves similar to the work on biological networks by Pan et al., who note the inability of a network to capture all known interactions (recall of 1) but still allows for a comparative analysis between datasets or networks [48]. However, a direct comparison of values is more difficult when network sizes differ, since the number of nodes influences the total number of actual positives. Nonetheless, overall trends can be insightful and compared between networks. The threshold for when an interaction is considered true can be varied between STRING interaction scores of 0 and 1000, which means that for each of those values a value for precision and recall can be calculated and plotted. The precision-recall analyses in this work do highlight STRING interactions with a score of 400, indicating medium confidence in the evidence for an interaction.

CORUM4.0 Validation

The CORUM database [49] is a widely-used collection of manually curated mammalian protein complexes to validate the ability of network interactions to reflect known biology [13], [28], [50]. The database, started in 2007, has received frequent updates with the latest 4.0 version being released in January 2023. At time of writing, the CORUM4.0 database servers are suffering from a cyber-attack and are therefore not available for a public download. Fortunately, the CORUM4.0 complexes were sent to us by the database's author after an inquiry about the server outage. The CORUM team was hopeful that after the issues were resolved, the complex database could be downloaded again from http://mips.helmholtz-muenchen.de/corum/.

In a similar approach to Replogle et al. [13], we used a cluster-agnostic approach and looked for complexes who had at least 66% of genes present in the presented network. Nodes that had a degree of 0 were not considered. Since protein complexes are tightly connected, every complex was treated as a fully connected network with the genes in the complex as nodes and every node having an undirected link to all others in the complex. These links are used in the Results to show interaction strength of CORUM verified connections in the data versus those connections of the background distribution of connections between all possible gene pairs. In addition we can sum the number of edges that are part of unique CORUM complexes.

Gene Set Enrichment Scores

The curated databases used in this work all bring relevant aspects when evaluating biological relevance of resulting networks. To more uniformly compare the network's recapitulation of the databases, we perform an enrichment analysis in a similar fashion to Boyle, Pritchard and Greenleaf [50]. By only looking at gene-gene interactions, we determine co-annotation by any term in the GO, STRING or CORUM databases. An additional advantage is this approach's inclusion of a term for edges not present in the network but do show co-annotation in a database. This can be seen as a measure of the network's recall, because it penalizes the enrichment score when a network lacks many edges which it should contain if the network construction method and the dataset truly capture known interactions. The summed enrichment for these interactions are calculated as in Equation 2.3.

$$\texttt{Enrichment} = \log_2\left(\frac{E_{fc}/E_c}{E_f/E}\right) \tag{2.3}$$

Where E is the number of all possible edges between nodes in the network, E_c are the edges actually present in the network, E_{fc} are present edges which are co-annotated in a database, and E_f are edges not present in the network, but are annotated in an external database. To determine all possible edges between nodes in the network, we only considered nodes in the network with a degree > 0.

Control Cluster Behaviour

Prior to commencing this work, twelve groups of genes in the HAP1 dataset were curated by academics actively researching with the CELL-seq screening technology to find groups with highly similar phenotypic profiles amongst the genes. These groups were manually annotated and confirmed to reflect true biological processes and are confirmed to interact by literature. The group sizes range from three to five genes, and they are highly diverse in their representation of biological processes. As a control, we can track the behaviour of the genes belonging to the twelve groups and find whether each group individually clusters together in the network. Since the controls originate from the HAP1 compendium itself, their usage is heavily specific for the HAP1 dataset only. The twelve groups of genes and their associated GO:BP terms are found in Table B.2.

The twelve groups showed quite different interaction strengths in DepMap however, where some groups translated almost one-to-one, some other groups did not contain any strong inner-group connection. For example, the group associated to the endosomal transport GO:BP term was also present in DepMap with very high interaction strengths, while the protein O-linked glycosylation group did not show any meaningful interactions in DepMap, since the genes in that group are related to muscle development and do not necessarily score in synthetic lethality screens. As such a different set of controls is required for DepMap.

The 2021 work by Shimada et al. introduced the ECHODOTS clustering algorithm to find clusters of genes with similar dependency strengths that are robust to stochastic influences from the t-SNE algorithm [29]. They applied the ECHODOTS algorithm to DepMap, retrieving clusters with highly correlating profiles across cell lines and created a web-tool to explore them dubbed shinyDepMap[2]. Using this tool, we recovered ten groups of genes that are similarly sized to the HAP1 control groups, and were further required to each be enriched for a GO:BP term. To aid comparison, seven of the groups recovered are related to one other HAP1 control group through an overlap in genes and similar GO:BP term enrichment. An additional three groups were added based on them being in the top scoring groups from the ECHODOTS algorithm listed in Figure 5B of the shinyDepMap work. The control groups are listed in Table B.3.

2.3.3. Network Metrics

For each generated network, we examined both connectivity and functional metrics to gain a more complete understanding. Connectivity metrics, like node count and degree, gauge the number of nodes sharing similar interaction strengths among nodes. We also tracked the number and sizes of connected components, which are linked connected subgraphs separate from other nodes. The count of components gave insights into network fragmentation as interactions changed, impacting functionality. For instance, stricter interaction strengths could isolate previously connected components, disrupting information flow. As an additional reference, Supplementary Table B.1 contains the brief definitions for characteristics of networks that are frequently referred to in this work. The characteristics; degree, cluster coefficient, betweenness centrality and neighborhood connectivity, are common in literature and their definitions listed here are from Barabási's book on network science [51].

For functional metrics we used two approaches. The Leiden algorithm was applied to detect communities of highly inter-connected groups of nodes. We can thus track the number and sizes of communities found. Edges in these communities as well as in the overall network were validated according to the approaches in Section 2.3.2. The second functional metric is to track the behaviour of groups of genes which are well-established in the literature to interact and are captured by the HAP1 screen compendium with high interaction strengths (see Section 2.3.2). We know a priori that a network which accurately reflects biological processes should have the genes in each group in the same community, and most preferably those communities should include only one gene group each since the groups have high interaction strengths and are related to diverse and specific GO:BP terms. They therefore act as positive controls for the community detection.

2.3.4. Visualization

In general, graphs were treated as `Graph` objects from the `igraph` Python library [52]. As is common practice for biological networks, results were visualized in Cytoscape 3.10 [53]. Each created network's layout was produced with the edge-weighted spring embedded layout in Cytoscape, similar to SAFE [40]. For larger components consisting of many interactions, the prefuse force directed layout was applied as it more quickly calculates a layout and so it can be efficiency applied to the giant components of networks presented in this work. Nodes are coloured according to the biological process the found community belongs to. Nodes with a bold white line are evidence genes, meaning that they are annotated by the Gene Ontology to the GO:BP term for which the entire community

[2] https://labsyspharm.shinyapps.io/depmap/

is enriched. Edges between nodes range in the amount of whitespace or "dashedness" from a straight line, to a line with points and dashes, to a dashed line, and finally a line with only \characters. These four categories reflects the categories for strength of interactions in the STRING database. The more whitespace a line contains, the less strongly it is connected in the STRING database. Finally, diamond shaped nodes are those part of CORUM4.0 annotated complexes.

2.4. BACKGROUND ON METHODS

2.4.1. CLUSTERONE MODULE DISCOVERY

Nepusz, Yu and Paccanaro introduced the ClusterOne algorithm in 2012 as a density-based clustering approach with the ability to find overlapping clusters [54]. The philosophy behind the algorithm is to identify groups of densely connected nodes by measuring their cohesiveness. Cohesiveness is a quality function that uses the assumption that well-defined regions in the feature space have many within-region connections and few boundary interactions, meaning interactions between nodes within the region and nodes outside the region. The algorithm starts with one seed node as a found cluster V. In a greedy fashion, the algorithm adds another node to this cluster which increases the quality function of the cluster the most. The quality (cohesiveness) of V, is defined in a similar fashion to [54] in Equation 2.4.

$$f(V) = \frac{\Sigma e_{in}}{\Sigma e_{in} + \Sigma e_{bound}} \tag{2.4}$$

e_{in} are the edges within the cluster V, while e_{bound} are those edges between nodes in V and nodes outside of V. In this work, the edges are weighted and therefore the cohesiveness function sums over the weights of e_{in} and e_{bound}. Nodes can also be removed from V at any point if that improves cohesiveness. This process continues until no neighbors of any vertices in V can be added that improve cohesiveness.

As the algorithm performs this approach for every node, often nodes are part of multiple clusters. Those clusters with a high overlap are merged together to form one cluster. In the final step, the density of all found clusters is compared to a user-defined threshold d, and only those with a higher density are kept in the final results. For the clusters, their density is calculated as $d(V) = \frac{V_e}{V_n(V_n-1)/2}$, where V_e is the number of edges in V and V_n is the number of vertices in V. It is thus a ratio between the number of edges in V against the maximum number of possible edges. Weinberg et al. [28] logically relate this parameter d to cluster sizes, as clusters with fewer nodes need relatively fewer within-cluster connections to produce a higher density $d(V)$.

2.4.2. ARACNE-AP NETWORK CONSTRUCTION

One of the algorithms applied to assess whether $\mathbf{X_H}$ was out-of-the-box suited for analysis by known gene-gene network inference algorithms is ARACNE-AP [55]. ARACNE-AP uses a different approach to traditional correlation-based relationships between samples in a dataset, instead employing mutual information I to measure the dependence between two variables [56]. The most interpretable definition of mutual information is given in Equation 2.5, where D_{KL} is the Kullback-Leibler divergence. Intuitively, the mutual information between genes X and Y ($I(X,Y)$) is 0 when X and Y are independent, meaning that observing X will not provide any information about Y. As a property of D_{KL}, the mutual information is non-negative and has no upper-bound.

$$I(X,Y) = D_{KL}(P_{(X,Y)}||P_X \times P_Y) \tag{2.5}$$

ARACNE requires a single run to estimate the threshold for Mutual Information, to estimate when interactions are discarded from the network. Then 100 runs of the algorithm with bootstrapping were ran, which is the same amount of runs as the authors of ARACNE-AP use on The Cancer Genome Atlas dataset. Then all 100 runs are combined into a single consensus network, based on the statistical significance of the number of times an edge is found in each bootstrap, using Bonferroni corrected P values < 0.05 as threshold. Technical details for reproducibility are found in Supplementary C.

2.4.3. MUTUAL K-NEAREST NEIGHBOR NETWORKS

In this work, we define the MKNN graph in a similar fashion to Zhang, Kiranyaz and Gabbouj [57]. **D** was previously defined as a symmetric distance matrix containing the GLS P value between every pair of genes in either the HAP1 screen compendium ($\mathbf{D_H}$) or the DepMap dataset ($\mathbf{D_D}$). Let d_i represent one row of such a distance matrix for a gene i, such that d_i is the vector of GLS P values between a gene i and all other genes in the dataset. Then let $K(d_i)$ be the set of k nearest neighbors of data point d_i, with obvious connotation that $|K(d_i)| = k$. Again let $G = (V,E)$ be a graph with every gene being represented as a node $v \in V$. We connect a pair of nodes v_i and v_j if the two end data points d_i and d_j are in each other's set K. Formally, $(v_i, v_j) \in E \implies d_j \in K(d_i) \wedge d_i \in K(d_j)$. The constructed graph is the MKNN graph, note that it is undirected and variable depending on the chosen value for k. The MKNN graph can be stated to be more stringent, arising from the reciprocal affinity criteria, and can be considered a subgraph of a KNN graph. In a KNN graph, every node's degree is always $\geq k$, while in the MKKN graph every node's degree is always $\leq k$.

2.4.4. DISTRIBUTION FITTING

We employed Scipy's `scipy.stats.rv_continuous.fit` function to robustly estimate parameters for a distribution model, where the data entered are the values from **D**. This was done individually for every gene in the dataset. We employ the Kolmogorov-Smirnov (KS) test as a statistical tool to estimate goodness-of-fit between the observed data and the estimated distribution. We perform a one-sample KS test, where a distance is quantified between the empirical distribution function of the sample and the cumulative distribution function (CDF) of the generalized Pareto distribution (GPD). To determine the significance of the goodness-of-fit, we calculate the P value associated with the KS test statistic. The KS test is defined as such that when the P value is is high it means the null hypothesis of the fit being accurate is true and a low P value suggests the null hypothesis should be rejected. We employed the default $\alpha = 0.05$ threshold and defined all genes whose KS-test's associated P value was > 0.05 to be well-fit by the distribution.

2.4.5. Gene Groups for Separated Biology Control Networks

Section 3.4.3 covers the creation of networks from a set of 500 genes that are selected to be from five different groups of biology. One hundred genes each were selected that are related to either DNA repair, apoptosis, translation, transcription factors or the mitochondria. The genes annotated to these groups were gathered from highly curated sources widely used in literature. The mitochondrial genes are found in the MitoCarta3.0 collection [58]. The transcription factors are those from the 2018 work by Lambert et al. published in Cell [59]. For the remaining three groups, these were selected from hallmark human gene sets curated by the Molecular Signatures Database [60]. From those lists, we selected a sample of 100 from each such that every gene was present in both the HAP1 compendium and the DepMap dataset.

2.4.6. Biological Coverage Analysis: Obtaining Broad GO:BP Terms

The GO:BP database is organized as Directed Acyclic Graph (DAG) data-structure, where the general Biological Process GO term is situated at the root and GO terms are hierarchically annotated based on specificity of the process up until a depth of 11 where the leaf GO terms reside. One thing to keep in mind is that there can be multiple paths between two nodes, due to the intrinsic un-hierarchical ordering of biological processes in nature. GOATOOLS [61] is a library created for Python specifically to perform queries and enrichment analysis on the gene ontology (GO). The gene ontology was download from the official source at `http://geneontology.org/docs/download-ontology/`. For initial GO term comparison between datasets, the aim is to use the propagation of annotations property in the GO tree and for each term obtain the broader GO terms at the first and second levels of the ontology. To this end, we employed a breadth-first-search algorithm which started from a GO:BP term retrieved from the networks and move up the tree and located all the first and second level terms that could be reached from the starting term. All unique level 1 and 2 terms are stored, and as such one specific GO term can have multiple broader categories in the upper levels.

3 RESULTS

This chapter presents an overview of approaches used to establish networks that recover meaningful biological relationships in both the HAP1 and DepMap datasets. We further perform PCA analysis and comparative analysis to DepMap to investigate the contribution of the limited number of screens in the HAP1 screen compendium. After network and screen analysis, a comparative investigation on the coverage of biology of both datasets is performed. This brings forward hypotheses for potential new interactions the HAP1 screen compendium highlights. Finally, these hypotheses give suggestions on which phenotypes to cover with new screens.

3.1. DATASET ANALYSIS USING ESTABLISHED METHODS

To lay the groundwork for creating gene-gene interaction networks based on the HAP1 screen compendium $\mathbf{X_H}$, we establish a baseline by applying two known methods from literature to the dataset. This baseline gauges whether the uniquely defined HAP1 phenotypic profiles for genes shows straight out-of-the-box results for seminal methods used for gene-gene network inference, originally designed for different datasets. These results serve as a benchmark for future analyses. The experiments are conducted using the method proposed from the GLS method by Weinberg et al. [28] and the ARACNE-AP model [55]. We've opted for the GLS method since the work by Weinberg et al. validated the retrieved clusters in a network based on the DepMap dataset to recapitulate known biology by using an enrichment analysis similar to the approach employed in this work. In addition, the GLS preprocessing step of the screen compendium to a distance matrix was brought forward by their method. This approach can thus give insights into the HAP1 screen compendium when compared to DepMap. ARACNE-AP was chosen due its well-established presence in the literature [62] and its unique approach to calculating pairwise relationships between genes based on an information theoretic framework. As argued by Margolin et al. in the original ARACNE implementation [63], the information theoretic approach is more suited than traditional correlation-based approaches for capturing non-linear relationships between genes and also comparatively more robust to noise. ARACNE-AP was originally designed to work with a matrix of gene expression data, but mutual information is not exclusive to gene expression and can be applied to any two variables whose values can be modelled using a probability distribution.

ClusterONE

The effort of Weinberg et al. towards a genome-wide network can be summarized into three core contributions; the application of GLS to the dataset, finding clusters of genes with a clustering algorithm, and using diffusion maps combined with a UMAP projection for final mapping and visualization. Since the final mapping is mostly for visualization purposes and reliant on the results from the clustering algorithm, we confine our approach to only employ the clustering algorithm on the distance matrices **D**. The clustering applied is an established density-based algorithm termed ClusterONE [54]. ClusterONE finds potentially overlapping modules based on a network structure or distance matrix by greedily "growing" regions from nodes as starting points until the quality of the module is above a user-provided hyperparameter density *d*, which dictates how strong within-module interactions are relative to interactions to vertices outside the module. A background on ClusterONE and its parameters are further described in Section 2.4.1.

As the functioning of this algorithm is highly dependent on the value of *d*, [28] uses the values 0.2, 0.5 and 0.9 for *d* and treats the resulting modules as a singular output from the algorithm. For weighted graphs, the default value is 0.5. By using a range of values for *d*, multiple biological scales are captured, where low values for *d* return large clusters while high values for *d* correspond to smaller clusters of specific functional organization. We apply ClusterONE on the FDR corrected GLS P values in $\mathbf{D_H}$ and $\mathbf{D_D}$. After ClusterONE analysis, quality of the resulting clusters are assessed by performing GO:BP Enrichment Analysis, as described in Section 2.3.2.

The number of resulting modules are shown in Table 3.1. Results for $\mathbf{D_D}$ are included as a control proving the implementation works as intended. Indeed for $\mathbf{D_D}$ we find a little over 200 more modules than originally found in [28], most likely due to the increased number of cell lines included in the newest iteration of DepMap used in this research. For $\mathbf{D_H}$, we see the algorithm is not suited for finding larger-scale clusters, with the $d = 0.2$ adding every node and edge to a single cluster after 10 days of runtime. After seven days of runtime on the NKI's RHPC cluster, 14% progress was reached using $d = 0.5$, therefore ClusterONE using $d = 0.5$ on $\mathbf{D_H}$ is considered not feasible. The program was still running, indicating they most likely got "stuck" in calculating a module which never reached the exit condition of its density. Since $d = 0.2$ returned one cluster

	Modules in $\mathbf{D_H}$ (HAP1)	Modules in $\mathbf{D_D}$ (DepMap)
density $d = 0.2$	1	367
density $d = 0.5$	N.A.	1995
density $d = 0.9$	612	3077
Total	**613**	**5439**
Total % of genes in modules	15.3%	88.1%

Table 3.1: Clusters generated by ClusterONE [54] on both datasets following the procedure described by [28].

after ten days, most likely there will be similarly spaced intervals returning large modules for $d = 0.5$. We hypothesize that since the bulk of interactions have a median GLS P value between 0.4 and 0.6 (see Figure 2.3), many edges are added without it being detrimental to the total scoring of within-group edges relative to the boundary edges, and thus the exit condition requires a stricter quality parameter.

For $d = 0.9$, 612 clusters were generated with GO:BP enrichment analysis showing 439 are enriched for a GO:BP term, with 211 distinct terms in total. The high number of overlapping GO terms is due to the fact that genes can be constituents of multiple clusters, which leads to many clusters with a relatively large overlap being enriched for the same term. This result indicates that $\mathbf{X_H}$ is able to capture known biological processes with relatively small-scale clusters, ranging from 700 genes to 5 genes. Processes with the highest enrichment P values ($-\log_{10}$(P value)) pertain to mitochondrial gene expression (185.2), post-translational modification of proteins (42.3) and organelle membrane transport (38.7). Biological processes like hippo signalling and TOR signalling are most prominent and account for 50 of the enriched communities. Table 3.1 does show that there is a large number of genes in HAP1 that are not accounted for in the modules. A lot of starting seed nodes thus did not form a module, the following sections that do not use ClusterONE will further analyze this behaviour of $\mathbf{D_H}$.

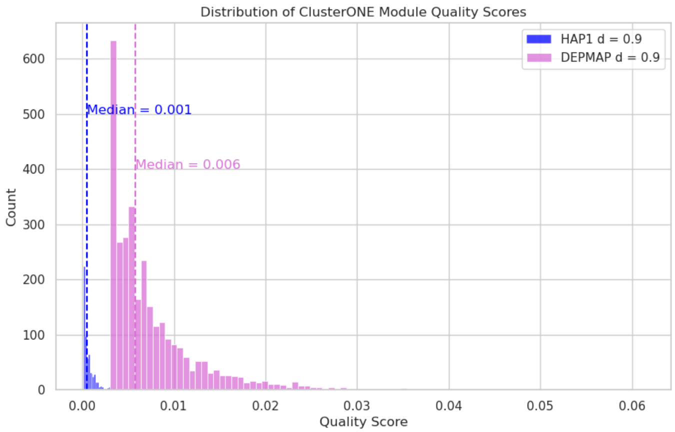


Figure 3.1: Distribution of the ClusterONE module qualities for $d = 0.9$.

The distribution of ClusterONE's quality function, as defined in the Methodology, is shown in Figure 3.1. Both the the spread and overall quality of modules are lower in $\mathbf{D_H}$ than in $\mathbf{D_D}$. Another important metric for the modules is the depth of the GO terms for which they are enriched. A greater depth indicates a greater specificity of the biological process. The average GO Term depth of ClusterONE terms is 5.37, where the maximum attainable depth is 11 in the GO:BP tree. Comparisons against this number will be made in following approaches. In conclusion, ClusterONE does capture groups of genes significantly enriched for a biological process, but there is room for improvement on the HAP1 screen compendium. Its failure to provide meaningful clusters on low densities indicate that only relatively smaller groups of nodes show coherent clustering, likely due to the

MC sampling already filtering for strong interactions to be retained in the dataset, or a lacking number of screens in the compendium. However, dense clusters are found to recapitulate known biology in smaller groups of genes albeit with less coverage of genes and number of modules than DepMap. These result spark several investigations in subsequent sections.

ARACNE-AP

The ARACNE-AP algorithm uses an information theoretic approach along with bootstrapping of the samples to provide a network more robust to noisy data and able to capture non-linear relationships. The background for ARACNE-AP is given in Section 2.4.2. ARACNE-AP was applied to both the original $\mathbf{X}_\mathbf{H}$ and the distance matrix $\mathbf{D}_\mathbf{H}$. Experiments on $\mathbf{D}_\mathbf{H}$ were not successful, as the matrix of 19385 x 19385 required too much memory for even a single bootstrap, which was not feasible even on the NKI's RHPC cluster. Application on the original $\mathbf{X}_\mathbf{H}$ should serve as an indicator for the success of non-linear methods on the raw dataset, unaltered for issues relating to screen non-independence and biased linear correlations. The ARACNE-AP retrieved network is shown in Figure 3.2.

We see a clear distinction of two densely connected components, sparsely connected by a few intermediary genes. The largest component, consisting of 15540 nodes, is enriched for being a cellular process with a $(-\log_{10}(\texttt{P value}))$ of 81.2, which is a very broad GO:BP term for which 14266 genes are annotated. The smaller green component is enriched for the G protein-coupled receptor signaling pathway (17.6). While not as broad as cellular process, the G protein family acts as molecular switches in cells and are involved in GTP/GDP signalling, which is an essential mechanism and not limited to a single process. A remaining large group of 712 genes is enriched for the sensory perception of smell (20.5). This shows how GO:BP enrichment analysis is not always insightful for such large terms as sensory perception of smell.654 genes in that group are not associated with that term, so the term's precision is low at 0.08, but since we do capture 57% of the term's annotated genes in the 712, it is enriched.

The ARACNE-AP approach does cover all genes in comparison to ClusterONE, but

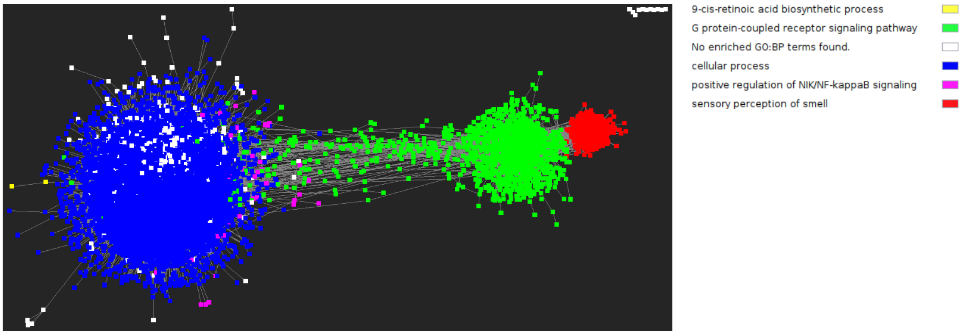

9-cis-retinoic acid biosynthetic process 🟨
G protein-coupled receptor signaling pathway 🟩
No enriched GO:BP terms found. □
cellular process 🟦
positive regulation of NIK/NF-kappaB signaling 🟪
sensory perception of smell 🟥

Figure 3.2: The interaction network derived from ARACNE-AP applied on $\mathbf{X}_\mathbf{H}$. Genes are represented as nodes, edges are mutual information values which are above the pre-calculated threshold, and nodes are coloured by enriched GO:BP term.

is not suited for finding communities that are meaningful in a biological context. The GO:BP terms covered are very broad and do not provide much information on gene-gene interactions. Calculating mutual information involves estimating a probability distribution for each gene, and such a task becomes quite difficult with the low number of samples (167 screens). As these samples are also not corrected for not scoring in screens and nonindependence of screens, these distributions can also become quite related, and even the applied bootstrapping procedure does not account for the relatedness of most genes. As their mutual information values become so similar, it is harder for any clustering algorithm to determine clusters of smaller size.

In conclusion, out-of-the-box application of established methods for network inference do provide a baseline to compare further analyses against, but are not showing overall satisfactory results when applied to the HAP1 screen compendium. Found clusters are either too broad to provide meaningful insights or cannot be found with ClusterONE in the HAP1 screen compendium. However ClusterONE does find enriched modules that do capture biological processes, but compared to DepMap there is a fewer amount and they are of lower quality according to the ClusterONE's quality metric. The dataset does either not have enough features to be informative, or the phenotypic similarities of the genes do not lend themselves perfectly to these methods, which were originally designed for PPI datasets for ClusterONE and microarray expression profiles for ARACNE. In those cases, PPI datasets are more suited for ClusterONE since often they are composed of several experiments that do not necessarily cover the same parts of the interactome (diverging cell lines or experimental design) and proteins often form complexes with similar interaction profiles, both reasons can lead to denser more modular connections in PPI profile matrices [64]. Microarray data in general measures the expression levels of thousands of genes simultaneously, and as such the data matrices contain many features, leading to works only selecting those features that are most varied across the samples. This abundance of features is much better suited to ARACNE-AP's information theoretic approach, where a probability distribution needs to be modelled for each sample.

3.2. INTERPRETABLE NETWORKS TO DETERMINE MEANINGFUL GENE-GENE INTERACTIONS

We next sought to investigate the underlying cause for the varying results following analyses by known methods. Since there is no background information on the structural features of the HAP1 dataset's distance matrix $\mathbf{D_H}$, this requires a more fundamental approach. Therefore, to more systematically determine the networking and clustering behaviour of our dataset, we apply two interpretable and commonly used methods for pairwise relationship modelling in a graph. The methods are i) networks with nodes that are connected only when edges are stronger than a global cut-off value and ii) a k-nearest neighbor (KNN) network.

Global interaction threshold network

Let $G = (V, E)$ be a graph. To determine whether two vertices $v_1, v_2 \in V$ share an edge, the global interaction cut-off graph is defined by function f_{global} shown in Equation 3.1.

$$f_{global}(v_i, v_j, \alpha) = \begin{cases} 1 & \text{if } \mathbf{D}_{i,j} \geq \alpha \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

Two vertices in G are connected only if their value in the distance matrix $\mathbf{D}$ is above a predefined threshold α. Note that the values in the distance matrices were $(-\log_{10}(\text{P value}))$ scaled such that higher values equate to more significant GLS P values. Global interaction cut-offs are frequently used in literature to determine when to connect nodes, such as in the seminal global yeast interaction network from the Boone lab [9] or when mapping genetic interactions in cancer cells as done by Rauscher et al. [11]. There is no standard value for α as it is dependent on the data and research objective and Rauscher et al. discussed the trade-off between the amount of information added into the network (high α) versus its usefulness in analysis (low α). We therefore conduct an analysis of network behaviour in a more systematic manner by measuring several network properties (Section 2.3.3) over a wide array of values for α. We can determine how stable the network is to reduced strictness of interaction threshold [28].

Initial results of the analysis while varying from P values of 0.01 to 0.5 in 50 steps is shown in Figure 3.3. Interestingly, we see a clear staircase pattern in the number of nodes, showing that nodes are clearly grouped by interaction strengths at certain thresholds, since the number of nodes also stays equal for a few values of α. This is due to an important feature of the data which has to be mentioned. The results of the MC sampling procedure (see Methodology), which was used to create $\mathbf{D_H}$, did not precisely contain the median GLS P value from the 1000 sampled GLS matrices. Instead the resulting P values were discretized into 400 equi-sized bins, which led to a maximum of 400 unique values being used in the $\mathbf{D_H}$ distance matrix. As such, the jumps in the number of connected nodes can correspond to the α cut-off edging over one of the 400 binned GLS P values, thereby at once including all edges that correspond to that bin. This discretization was performed prior to the commencement of this thesis work as a method to conserve memory during the memory-intensive sampling step. The analyses performed specifically in this section 3.2 are fully based on this discretized version of $\mathbf{D_H}$. Later in this work, we see that this discretized $\mathbf{D_H}$ was not suited for certain analyses, which will be discussed appropriately [1].

The leaps in number of nodes are thus explained by the discretization, but the overall trend is indicative of another phenomenon. Almost all nodes were part of a single large connected component as shown by the overlapping red and blue lines in Figure 3.3. We observe that already at P values < 0.05 there is one large network component while the remaining nodes remain either fully disconnected or forming pairs of nodes. The network representative of f_{global} with $\alpha = 0.01$ (the first network in Figure 3.3) is shown in Figure 3.4. The Leiden clustering algorithm is unable to distinguish smaller communities in the network, which is densely connected due to the high number of edges relative

[1] Note that section 3.1 does not use the discretized version as these analyses were re-performed for a more complete analysis later in the project's timeline.

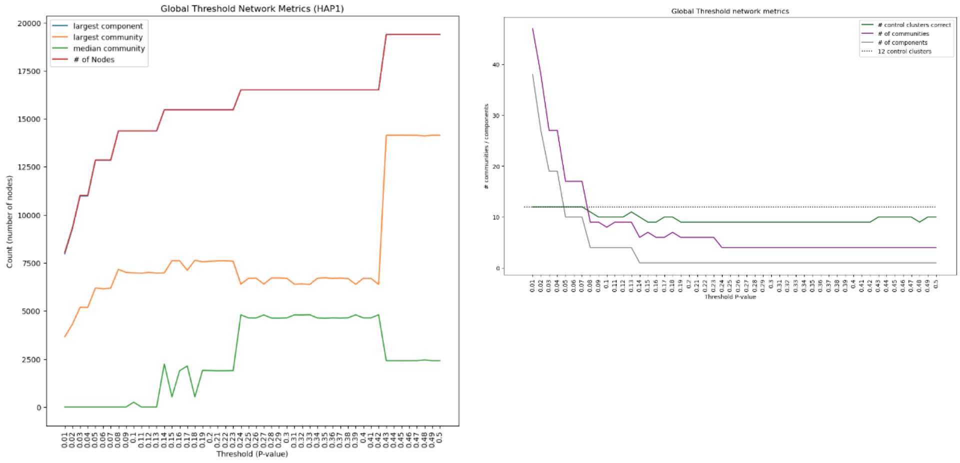


Figure 3.3: Analysis of connective and functional metrics for a range of α values from 0.01 to 0.5. The left graph tracks global metrics relating to number of nodes and component sizes. The right graph shows results of community finding.

to the number of nodes. When looking at the number of positive control clusters correctly clustered for this network in Figure 3.3b, all twelve groups are clustered together. They are however all found within the same community, while these twelve groups are functionally very different, indicating the lack of information such a global network retrieves in this setting. Figure 3.3b shows the number of disconnected components over the range of values as well as the number of communities found by the Leiden algorithm, showing they are heavily related. Disconnected components, found by Leiden to be distinct communities, can no longer be distinguished as seperate commmunities when the components merge.

To gain further insights into the topological organization of the network, we analyzed the variation of nearest neighbors given by the degree distribution [65], shown in Figure A.1. The degree distribution provides valuable insights into how nodes are connected and how information or influence flows within the network. The degree distribution of the network appears to follow a power law, where a small number of nodes referred to as *hub* nodes have a high degree and connect the remaining nodes with sparser degrees, which is a well-documented feature of biological networks [66]. However, two important features of the distribution do not conform to power law behaviour.

First, there are a very small number of nodes with an incredibly high degree, making the spread of the distribution extremely broad. These hub nodes are not completely unmerited. The genes with highest degrees on average score in a high number (> 50) of screens. The top 5 hub nodes by degree (degree varying from 1778 to 1710) are NRXN1, DLG2, TRPM3, ROBO2 and SYT1, which also make sense in a biological context. NRXN1 encodes a membrane protein involved in cell-cell-interactions, and required for the formation of complexes for efficient neurotransmission and the formation of synaptic contacts. DLG2 encodes a protein involved in scaffolding for receptors ans signalling proteins, and is known to be related to the RAF/MAP kinase cascade and unblocking of

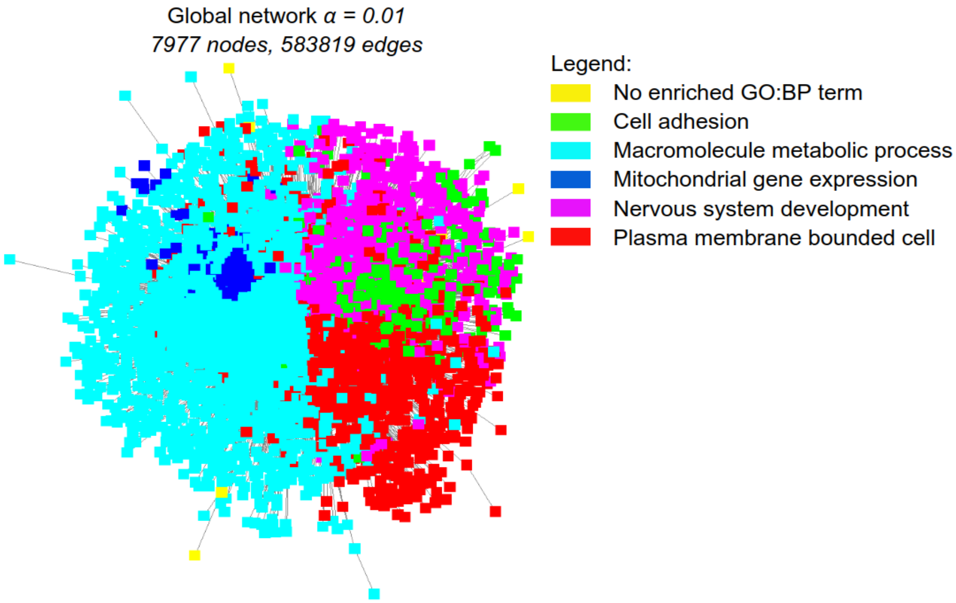

Global network α = 0.01
7977 nodes, 583819 edges
Legend:
🟨 No enriched GO:BP term
🟩 Cell adhesion
🟦 Macromolecule metabolic process
🟦 Mitochondrial gene expression
🟪 Nervous system development
🟥 Plasma membrane bounded cell

Figure 3.4: f_{global} network using $\alpha = 0.01$. Leiden community algorithm returns five large clusters related to broad GO:BP terms.

NMDA receptors. TRPM3 encodes a protein belonging to the transient receptor potential family, responsible for cellular calcium signaling. ROBO2 encodes a protein highly conserved between human and fly's, and is a transmembrane receptor functioning in axon guidance and cell migration. SYT1 encodes a member of the synaptotagmin protein family, another integral membrane protein family that play a role in membrane signalling during trafficking of synaptic vesicles. The hub nodes are all essential for signalling processes, especially related to neural regions, and as such their high involvement with many processes involving a number of other genes is not unwarranted.

Another important characteristics of these genes is that they are relatively long in terms of base-pairs (bp) on the genome. The lengths are retrieved from BioMart [67], a well-established tool used for analyses on gene lengths like in the work by Lopes et al [68]. As the CELL-seq screens are performed with randomized gene-trap procedures, larger genes simply have a higher probability of being truncated than smaller genes. As such, they have a higher number of insertions when compared to the other genes. The method to calculate the MI of genes as described in the dataset description in the Methodology uses the number of in-gene insertions. When the values for in-gene insertions are relatively higher when compared to the out-of-gene insertions, the MI values tend to become larger too, and also influences their P values for significance in screens. The hub nodes truly reflecting a biology or being a result of the experimental approach (or both) is a point to investigate further in the following sections.

A second interesting feature of the distribution is a very large concentration of nodes with a degree of 1, which is not conform the power-law. As such the distribution appears bi-modal. This can be indicative of a commonly observed phenom in (biological) networks called a core-periphery structure. The 1999 work by Borgatti and Everett first formalizes this structure as a meso-scale feature of networks where an undirected network can be broadly partitioned into two sets, where one set contains a densely connected *core* which is surrounded by a *periphery* of nodes which have sparse connections within the periphery itself, but have more connections to the core [69]. This behaviour is visualized for the HAP1 compendium in the next section.

The core-periphery structure is extremely apparent in the HAP1 screen compendium when analyzing networks with an extremely low value for α. As larger values for α (> 0.01) showed similar behaviours, we opted to investigate lower values of α more extensively, and thus the range of α values is defined in log space to evaluate 100 values between 1E-10 to 1. The f_{global} network of $\alpha = 1 * 10^{-10}$ is shown in Supplementary Figure A.2. The P value threshold for interactions is the lowest investigated in this work. Characteristics of the network are summarized in Table 3.2. The strong interactions lead to a more favorable identification of 35 communities with enriched for a unique GO:BP term. By far the largest term is mitochondrial gene expression, almost completely covering the second largest component in the graph (light blue). This is unsurprising, as it was also one of the major clusters found with the ClusterONE algorithm. The network is in general quite sparse, consisting of multiple disconnected components. Of the 500 genes with the highest degree in the $f_{global} \alpha = 0.01$ graph, the most prominent hub nodes, 80% are found in this extremely sparse $\alpha = 1 * 10^{-10}$ network. This supports the conclusion that the more densely connected component in the $\alpha = 0.01$ network is largely due to the higher interactions strengths found between those nodes and for that reason are important in providing connectivity in the larger network.

Network Feature	**Value**
Number of Nodes	1521
Number of Edges	5670
GO:BP Enriched Communities	35
Total Communities	103
Giant Component Size	1041

Table 3.2: Characteristics of the f_{global} network with $\alpha = 1 * 10^{-10}$

The separation of a dense and strongly connected core overlapped with a much larger and sparser periphery component apparent in the network is an actively researched topic in statistics [70]. However, finding a suitable approach for this dataset would not contribute to answering the research questions. A strongly defined "core" would not give any information on the remaining genes, while one of the crucial ideas behind the HAP1 screen compendium is that gene-gene interactions that are associated by more specific phenotypes should cover areas of biology less prominent in datasets that measure broad phenotypes such as survival. Furthermore, a threshold for the core would also have to be highly specific in regards to the biological processes involved. Defining one singular value for α to split the core does not cover the complex and diverse interactions in differing biological contexts. Not all pathways are created equally, and interaction strengths between genes in different pathways are not of singular strength throughout the cell and their measurements are heavily dependent on the experimental conditions. Setting a low value for α could thus lose information on pathways that are still very relevant. And trying to capture all interactions results in a noisy and uninterpretable network. The proposal of a core-periphery separation algorithm is therefore outside the score of this network. To potentially improve upon aforementioned disadvantages of one global threshold, we apply k-nearest neighbor (KNN) graph construction, which ensures all nodes are contributing to the overall network structure.

k-Nearest Neighbor Networks

k-nearest neighbor (KNN) approaches lie at the foundation of many non-parametric clustering algorithms, as well as biological networks like in the work by Weinberg et al. [28]. Additionally, prior to this thesis work the HAP1 screen compendium was analyzed using UMAP, which showed coherent clusters forming enriched for Reactome [33] terms, which is a database of biological pathways, and UMAP employs KNN networks for the dimensionality reduction [71]. These results are not included in this work, but the coherent clusters were only found with the number of neighbor parameters $k = 2$, the lowest possible value, meaning only the most local neighborhoods are able to recapitulate biology. UMAP specifically employs KNN networks since it is often applied to high-dimensional data, where distances between nodes tend to become larger but also more similar due to the curse of dimensionality. The difference between a node and its first neighbor could therefore be large, but the subsequent differences to other neighbors are smaller. This is an advantage of using KNN networks over a global interaction cut-off, as this adds a focus on local connectivity irrespective of absolute distances.

We performed a similar analysis to the global networks by analyzing the networks created by varying k over a range from 1 to 40. In addition to the structural and functional metrics, we track the total properties for a network to lend itself to community discovery and subsequent enrichment for GO:BP terms for those communities. Results for the grid search are presented in Figure 3.5. Presented results are showing k values from 1 to 10 to not overpopulate the graph, and in addition the number of communities in each subsequent network is similar to the $k = 10$ network.

Figure 3.5 shows the distribution of GO:BP enrichment P values for each community in the networks found with the Leiden algorithm for each value of k. The $k = 1$ network logically retains the largest number of separate components, and is shown in Supplementary Figure A.4. The largest component has 20% of total nodes and is split into 10 communities, while many other components are comprised of a singular community. From the Figure, 191 out of 429 communities found are enriched for a GO:BP term. This is in line with the findings from running the ClusterONE algorithm with a high density setting (small-scale clusters): there are biological processes captured by the dataset but clear separation of those processes requires sole consideration of a node's extremely local neighborhood of ($d = 0.9$ or $k = 1$), where we obtain the most enriched communities in absolute terms.

We can further evaluate the quality of the network using the STRING database and CORUM databases (Methodology 2.3.2), and plot the precision-recall curves of known STRING interactions in Figure 3.6A. The threshold of 400 are STRING interactions of

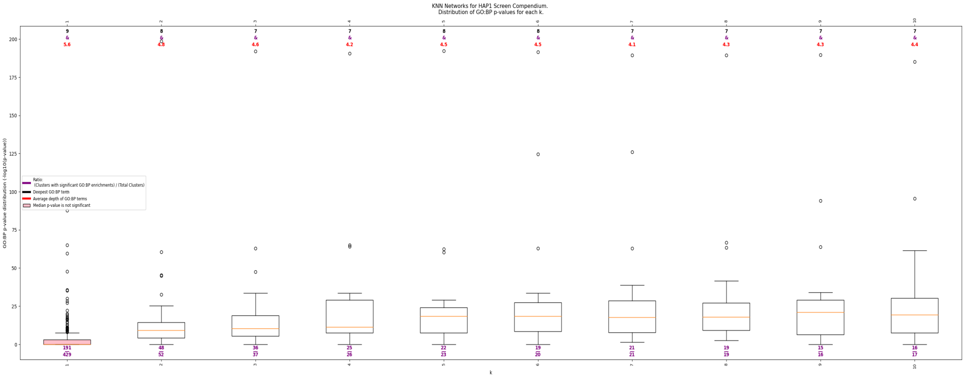


Figure 3.5: Distribution of GO:BP enrichment P values for communities in networks found over a range of k. Also noted are the deepest term found in the GO tree, as well as the average term depth.

medium confidence. The recall is very low, since all nodes are participating in the networks but the $k = 1$ ensures there are very few edges. The precision is comparatively high, slightly over half of edges are interactions in string with medium confidence. The enrichment in the CORUM database is shown in Figure 3.6B. The background distribution (blue) shows the distribution of GLS P values from all possible edges in the network ($V \times V$), which is similar to the observed GLS distribution from Figure 2.3. The red distribution are the GLS P values from the interactions in the network also found in CORUM. The interactions capture those relations in 3622 CORUM complexes out of the total 5204. This shows that the GLS P values of gene pairs found in CORUM are significantly stronger than the background distribution of all edges (by Mann-Whitney U test), but the distance between the medians is quite close. Furthermore, the low edge count of $k = 1$ leads to CORUM complexes being "hit" but not fully captured, as those are most frequently composed of 3 to 5 proteins many intra-complex interactions.

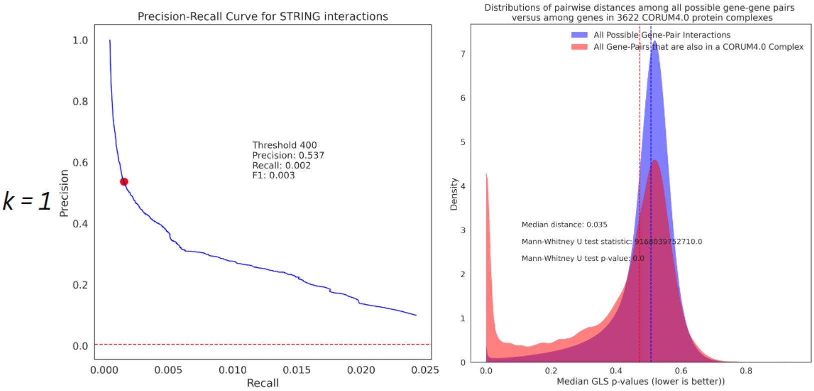


Figure 3.6: Precision recall curve for STRING and a GLS P value distribution comparison for CORUM4.0 of the KNN $k = 1$ networks.

When we increase k to 2, the entire network becomes immediately connected and forms one large component. The number of communities found reduces eight-fold, the Leiden algorithm struggles in finding small denser-connected regions in this connected component. As a result, community sizes increase, and while most of them are enriched we note that the statistical over-representation analysis which calculates enrichment is very dependent on the size of the community and the number of annotations to the GO term, meaning that as community sizes increase in general the GO terms for which they are enriched have more annotations. More annotations to a GO term are generally those associated to a broader category of biology, in other words specificity of biological processes covered is lost as k increases. This is further highlighted by the inability to identify the control clusters as distinct communities in all networks $k > 2$, while the $k = 1$ network does capture 9 out of 12 controls in separate enriched modules. For visualization, the $k = 4$ network is shown in Figure 3.7.

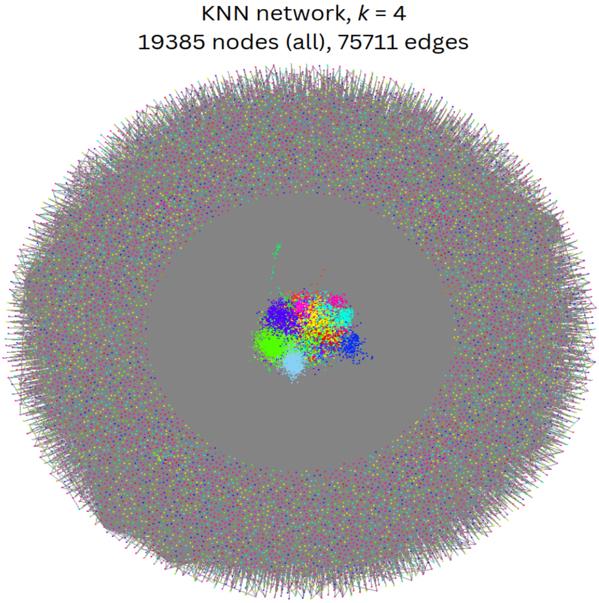


Figure 3.7: $k = 4$ KNN network. Nodes are colored according to community found.

This $k = 4$ networks visually functions as a representative of all $k >= 2$ networks. The KNN network structure has more clearly shown the core-periphery structure present in the data. We can naively separate the two partitions by simply selecting the nodes in the clearly visible donut shape as the periphery and the center as the core. The core contains 55.6% of nodes and 90.1% of edges, showing the densely connected core juxtapositioned with the sparse peripheral. This structure also gives insights into why the Leiden communities becoming increasingly large too, when examining the distribution of the neighborhood connectivity of nodes shown in Figure 3.8. Neighborhood connectivity of a node is the average degree of its neighbors, which has also been shown to commonly follow a power-law distribution in PPI networks [72]. In the case of Figure 3.8 there is a distinct bi-modality where one large group of nodes is very densely connected in its communities while another group has sparser connectivity. The Leiden algorithm has a resolution parameter that accounts for this, but this is pre-defined before analysis and is here more skewed towards larger communities if the amount of interactions make the clusters less modular.

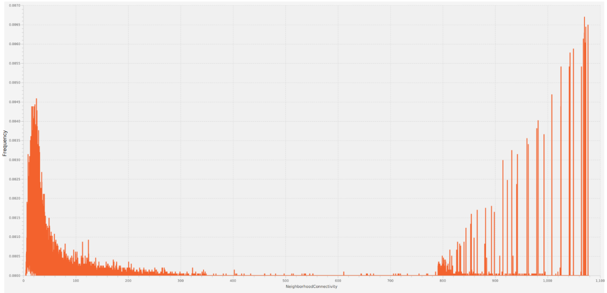


Figure 3.8: Distribution of neighborhood connectivity of the nodes in the $k = 4$ network

Such a structure does not indicate a direct flaw of the screen compendium; core-periphery structures are a feature of gene regulatory networks and PPI networks, where a set of master regulators or core proteins is highly conserved and performs essential functions, while the more peripheral regulation and proteins are related to more more organ-specific functions [73]. We asked whether this core-periphery structure was strictly a feature from biology or if is inherent to the HAP1 screen compendium. As noted in the dataset description, 11093 out of 19385 genes found in the screen compendium are a significant regulator according to their P values in at least one screen. The remaining are never significant, and have in common that they contain a lower number of insertions throughout all screens. We analyzed the proportion of genes in both partitions based on the amount of times they score as a regulator in a screen. The number of screens in which a gene scores for the $k = 4$ network is plotted in the pie-chart in Figure A.3. The periphery is dominated by genes that did not score in any screen, and 94% of genes score in either none or one screen. Additional distribution comparisons are illustrated in Figure 3.9, revealing a broader range of screen numbers within the core. Notably, genes with scores across up to 60 screens are also detected in the periphery, implying a lack of scoring is not the only reason for a gene to reside in the periphery.

Employing a Mann-Whitney U test, we find evidence to reject the null hypothesis that these samples share the same distribution. Hence, these distributions exhibit significant differences. We chose the Mann-Whitney U test over a conventional t-test due to the prevalence of genes that score in no screens, making the data not normally distributed. We note that the analysis shown here is for $k = 4$ only, however this trend continues for all values of k up to the analysis bound of 40. Those networks show an increasing shift of genes to this core structure until only 7 or 8 communities remain. Further analyses using the global and KNN networks using only genes found in significant screens resulted in more densely connected structures that did not lend themselves well to community analysis, akin to the "core-like" structures as above. This core-periphery relation is investigated more in future sections.

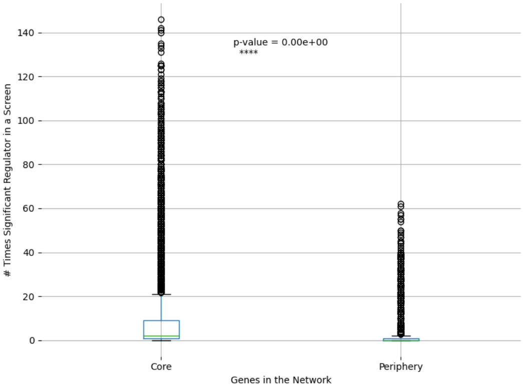


Figure 3.9: Data for the KNN network with $k = 4$. Comparison of the distribution of the number of screens per gene in which they are deemed significant. The Mann Whitney U test shows they differ significantly.

To summarize, the networks with more lenient global thresholds and k values >1 show a core-periphery structure, containing extremely large communities with broad GO:BP enrichment terms. The networks are too noisy and not able to effectively recapitulate biology. A global threshold is only suitable for the analysis of the dataset when it is so low that only the most extreme interactions remain. Likewise, only if $k = 1$ communities are found in the KNN networks which show enrichment for GO:BP terms and contain 9 out of 12 control clusters. We therefore continue to explore the dataset in order to bring forward the most biologically relevant and comprehensive gene-gene interaction network, knowing that local neighborhoods should be prioritized. We additionally investigate more thoroughly why the HAP1 screen compendium shows the above described network structures.

3.3. STRINGENCY OF MUTUAL K-NEAREST NEIGHBOR GRAPHS SHOWS MEANINGFUL CLUSTERING

To continue the investigation into local neighborhoods we propose the application of the mutual k-nearest neighbor (MKNN) graph structure on the screen compendium. This section argues why and how such networks address previous limitations and then continues with an analysis of the MKNN networks for a range of k values.

The MKNN approach introduces the concept of reciprocal affinity in the nearest neighbor relation: nodes are nearest neighbors if and only if both nodes are each other's k-nearest neighbors. While its use is less wide-spread than the standard KNN graphs, several variations for clustering algorithms have been introduced which exploit community structures found in MKNN graphs [57], [74]. In the work of Sardana and Bhatnagar, MKNN network structures are specifically employed to successfully find communities in synthetically created datasets that show clear core-periphery structures [75], which was clearly apparent in the previous network construction methods and limited the biological relevance of communities in the networks. Sardana and Bhatnagar also employ ClusterONE as a baseline and note similar results when applying their approach on a PPI network, which ClusterONE is designed for.

The 2021 work by Dalmia and Sia proposes the usage of MKNN networks in UMAP over the standard KNN structures [76]. They argue MKNN graphs can reduce the hub effect, highly relevant for our dataset problem in particular. The hub effect is the phenom that some nodes become highly connected hub nodes in the KNN graphs, which adds noise to the graph by losing local structure. Findings from the global and KNN graphs show the hub effect is present in the dataset: hub nodes found in the networks correspond to those nodes with overall highest interaction strengths and score the most in the CELL-seq screens and as a result dominate network connectivity patterns. In addition, the experimental design of CELL-seq screens biases towards longer genes being inserted into, leading to an increased presence as hub nodes shown in the global networks. In the KNN networks with $k \geq 2$ this is also apparent, some outlying nodes show a degree of over 3000 compared to the bulk of genes having a degree of k. Dalmia and Sia show a successful application of the MKNN structure that improves upon the hub effect and as such it has been implemented as an option in the official UMAP Python package. As such, the approach aims to correct for the biased hub nodes while not completely eliminating hub structures as they are present in biological networks. The hub node behaviour in MKNN networks is shown in this section and further analyzed in section 3.4. We provide a formal definition of MKNN graphs in the background section 2.4.3.

We again construct networks over a range of k values and show the GO:BP enrichment distributions in Figure 3.10. The MKNN graphs do not show immediate convergence to a smaller number of components like in the KNN $k \geq 2$ graphs. Instead they retain the feature of containing many separated components over the span of k values. This fragmentation of subgraphs is expected behaviour of MKNN graphs due to the added stringency, and can be interpreted as a positive attribute for the task of finding meaningful clusters, as reasoned by Ozaki et al. [77] who select the MKNN structure specifically for this purpose. The MKNN graphs indeed are more suited for finding clusters in this work as well, the fragmented components can be individually applied to the Leiden algorithm which struggled in finding communities in large subgraphs of the global and KNN networks. An example of the fragmentation is shown in Supplementary Figure A.5.

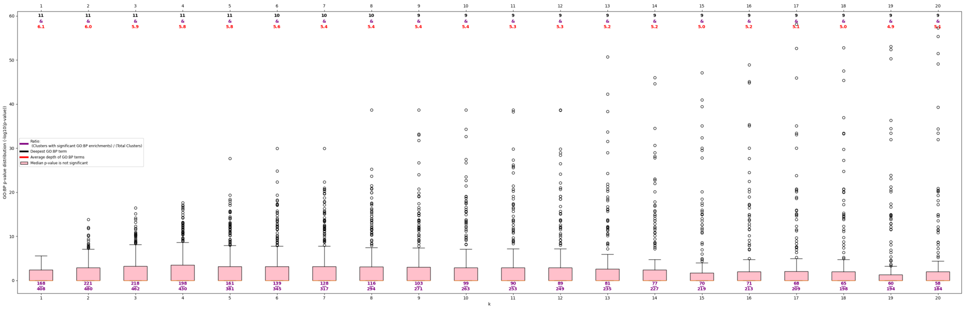


Figure 3.10: HAP1: Distribution of GO:BP enrichment P values for communities in networks found over a range of k. Also noted are the deepest term found in the GO tree, as well as the average term depth.

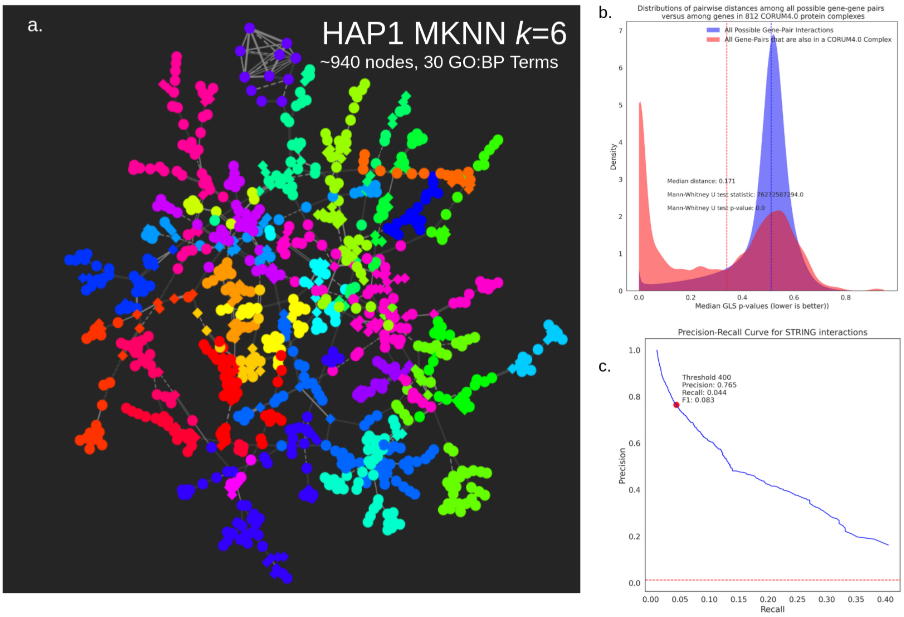


Figure 3.11: A: Largest subgraph of the MKNN $k = 6$ network, nodes colored by unique GO:BP term. B: Shows that the GLS P values of edges found in the network (red) that are associated to the same CORUM complex are in general much lower (=better) than the P values of all possible edges. C: Enrichment of edges in the network for all three GO categroies, STRING, STRING with intermediate strengths and higher, and CORUM.

As k increases, the most notable change is the increase in size of the largest component, as the smaller components merge together. In the KNN networks, this large component, there observed to be "core-like", has the downside of being too dense for clustering algorithms to find meaningful specific clusters. Due to the added stringency of MKNN however, the smaller components merge with relatively few new connections, and the large component retains separability into Leiden clusters. This can be seen in the giant component of the $k = 6$ network in Figure 3.11. The large component consists of around 940 nodes, which were determined by the Leiden algorithm to consists of 30 relatively dense communities, each one enriched for a GO:BP term. The separability is evident by the presence of 11 out of 12 control clusters being identified as communities, each enriched for a GO:BP term related to those listed in Table B.2. Not only do we find more controls than in the KNN $k = 1$ network, they are connected into one component too, which is beneficial because the network should ideally highlight all the interaction patterns in a cell. The remaining network consists of smaller subgraphs usually containing 1-3 enriched communities, and a considerable group of nodes that do not connect to the remaining network, which is an observed phenom in MKNN networks [76]. The behaviour of these isolated vertices is addressed in Section 3.4.

To assess the capability of the network to recall known biology, we compared edges in the network to known manually curated databases of biological interactions. Figure 3.11b shows again a CORUM validation, and compared to the KNN $k = 1$ in Figure 3.6 the medians are further apart and the peak of GLS P values with high interaction strengths is more pronounced. Figure 3.11c shows the precision-recall curve, which at 400 threshold are better than the KNN $k = 1$ network. Because precision and recall as presented here are dependent on the number of nodes and edges in the network, these values are higher due to the many non-connected nodes in the MKNN networks, making the comparison more unfair (see Methodology 2.3.2). However, in the KNN networks the $k = 1$ is the only network showing any result, and precision and recall drop-off steeply with the increase of k. In the MKNN these values are much more stable as k increases, however there is a trade-off between precision and recall: we capture more edges as k increases but they become increasingly less informative when comparing to biological databases.

The MKNN method retains a large amount of enriched communities over the span of tested k values. The largest number of enriched terms found in $k = 2$ with 221 communities enriched for a GO:BP term. 205 of those GO:BP terms are unique, a similar amount to the 211 unique GO:BP terms found by ClusterONE (Table 3.1 has 612 clusters for HAP1 with 211 unique terms). As mentioned in the ClusterONE analysis, average GO Term depth of ClusterONE terms is 5.37, versus 5.91 in the $k = 2$ network and 5.59 in all k networks (deeper indicating more specific categories in the ontology).

We compare the results of the MKNN on the HAP1 screens to MKNN networks constructed from DepMap, varying again over a range of 1 to 20 k values. The GO:BP enrichment distribution is shown in Figure 3.12. Interestingly, in the lower values of k more communities are found in DepMap and more are enriched for a GO:BP term. The enrichment P values are also higher across the board compared to those in the HAP1 compendium networks. As k increases we find fewer communities, but the relative amount of enriched among those is quite high, indicated by the white box plots from $k = 6$ onwards to boast more than 50% of communities enriched. In terms of structure, the connectivity is much more apparent in this dataset than in HAP1. The largest component already contains around 60% of nodes at $k = 4$, and grows to become almost fully connected at $k = 20$, while in HAP1 this component remains relatively small. Most nodes in HAP1 are further singular and thus disconnected from all other nodes, while in DepMap this is a much less prominent feature and at $k = 20$ is barely apparent. In both HAP1 and DepMap however, we do report that the stronger interactions in terms of GLS P values are also those most found in the STRING database, showing that both datasets detect known interactions in their experimental approaches, as seen in Supplementary Figure A.6. As k increases, the MKNN remains stringent on HAP1 and included primarily strong interactions, while at DepMap the number of edges and components becomes high and starts to include relatively lower GLS P values.

To conclude this section, the MKNN structures prove to be a more fruitful endeavour compared to the global threshold networks. The KNN network with $k = 1$ is more comparable in performance to the MKNN networks. However, the MKNN networks do capture more communities enriched for GO:BP terms, and as k increases this complete collapse to a core-periphery structure is not apparent. Instead, we find modular and informative components such as the giant component in the MKNN $k = 6$ networks. The stringency of MKNN networks keeps the focus on local connectivity, but becomes structurally much more akin to the desired gene-gene interaction network. The MKNN networks also capture a similar number of enriched communities to ClusterONE and

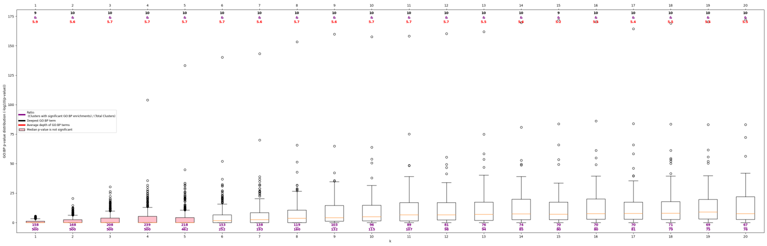


Figure 3.12: DepMap: Distribution of GO:BP enrichment P values for communities found in the networks over a range of k.

more control clusters than the $k = 1$ network.

However, there remain two flaws with this approach. First, the same k is used for all genes, and similar to the global and KNN networks the MKNN networks are therefore limited in their representation of actual biology, where some gene products act in many pathways compared to other genes related to more specific functions. Secondly, many genes are isolated vertices and do thus not add information to the network. That is a feature of MKNN networks and has the potential to reduce noise. However, we find that for example genes associated to the mitochondria are also widely spread throughout those unconnected nodes. In the global and KNN networks those genes were connecting much more frequently but their inclusion at more lenient thresholds turned the networks uninformative, and as such all proposed methods thus far do not bring forward a solid solution for this problem. It however indicates that the MKNN could be too stringent in its connection criteria. The next section continues with analyzing why and how these behavioral patterns of the MKNN networks are occurring, and whether this can be improved upon. Furthermore, the different connectivity patterns in HAP1 and DepMap should be investigated to answer more conclusively why they behave so differently. Finally, the knowledge behind genes' network behaviour can further contribute to making adjustments to this behaviour specific to the gene properties. Therefore the following section will additionally include approaches to include isolated vertices and further expands MKNN networks to incorporate gene characteristics.

3.4. How GLS Profiles Relate to Network Behaviour

The networks showed varying patterns in node connectivity and community alignment. Many of the nodes show a degree of 0, due to the stringent connectivity criteria in the MKNN network. Other nodes have a maximum degree of k and are part of large communities (> 50 nodes) and some form their own sub-network with fewer nodes. We asked what properties of the genes represented in the network determined their behaviour in the network. For this we made an inquiry into the properties of the distribution of GLS P values for each gene that denote connectivity to all other genes found in the distance matrix **D**. In addition, now that we have identified the MKNN networks to better mine the HAP1 screen compendium for relevant biological properties, we can continue with comparisons between the features of the HAP1 screen compendium to those of DepMap. As before, let $\mathbf{D}_H$ denote the GLS distance matrix for the HAP1 data and let $\mathbf{D}_D$ be the same for DepMap.

In order to extract meaningful characteristics from the distributions of GLS P values from each gene to all other genes, we tried to fit the most sensible statistical distribution on the distribution of each gene's GLS P values. The following approach to find the most sensible distribution makes no prior assumptions on the type or parameters of the distribution. Python's Scipy package [78] has circa 80 different distributions, and for a random sample of 500 genes we attempted to fit each of the distributions to the row in $\mathbf{D}$ belonging to that gene. How a fit was determined is provided in the background section of the Methodology 2.4.4. This showed not a single distribution fitted any of the 500 GLS P value distributions. The culprit was the discretization step performed on the MC sampling used to create the median GLS P values as was also found and described in the section on global threshold networks. The discrete nature of the GLS P values along with a non-uniform spread of the 400 possible values at each gene made fitting a distribution not possible. We therefore recalculated the MC sampled GLS distance matrices over a period of two weeks without a discretization step. Analysis shows that now all gene-gene GLS P values are now unique and expressed with 32-bit floating point precision.

Figure 3.13 A and C show the resampled distribution of GLS P values for a representative gene in both $\mathbf{D}_H$ and $\mathbf{D}_D$. The distribution in $\mathbf{D}_D$ is most easily interpretable, we see a Pareto distribution showing that most genes have low interaction strengths and relatively few genes are found with higher connection strengths. In the $\mathbf{D}_H$ dataset however, we see no such Pareto distribution, instead it is showing a consequence of the MC sampling as explained in Section 2.2 and shown in Figure 2.3. Again, attempting to describe the distribution with any of the circa 80 at disposal did not result in any gene being fit. However, the MKNN graphs have shown the best results for relatively low values of k, and also ensure that a node's degree is at most k. This means only the strongest connections are considered. This equates to the tails of the distributions shown in Figure 3.13. We argue that we can thus analyze only the tail behaviours of the distributions when the objective is to analyze the reason for behaviours shown in the MKNN graphs. When only analyzing the tails, we also see a more Pareto-like trend in both distributions for $\mathbf{D}_H$ and $\mathbf{D}_D$, making the comparisons between the two more equivalent.

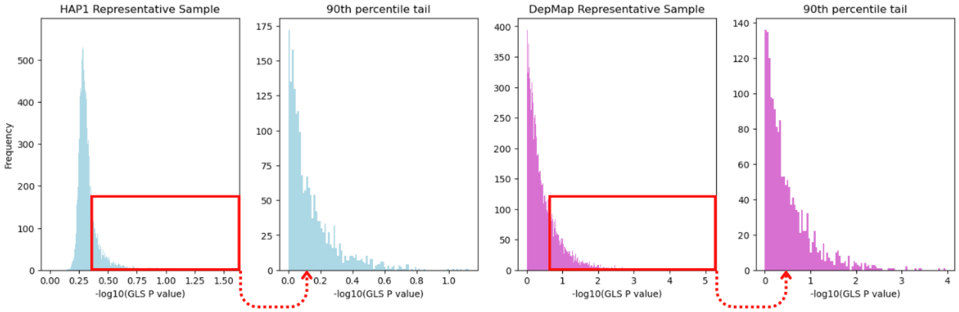


Figure 3.13: A and C: GLS P value distribution of a representative sample for $\mathbf{D_H}$ and $\mathbf{D_D}$. B and D: 90th percentile tail of the distributions.

A standard method to model the tail of another distribution is to describe the tail using a Generalized Pareto Distribution (GPD). We follow a similar method to the one used by Thijssen and Wessels [79] and use a rule of thumb that the tail of the distribution is estimated by 10% of samples. The GPD is described by three parameters, the location μ, the scale σ and shape ξ. μ corresponds to a location or threshold value above which the tail is modelled. σ controls the spread of the distribution and ξ determines the tail behaviour. The probability density function (PDF) of the GPD is given in Equation 3.2.

$$f(x;\mu,\sigma,\xi) = \frac{1}{\sigma}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}-1} \tag{3.2}$$

Let q denote the q-th percentile of the data for each gene, and so we set $q = 90$ for the 10% rule. Therefore, for each gene the value of q determines the location parameter μ. The visualization of the 90th percentile tails is shown in Figure 3.13 B and D. We then estimate the σ and ξ parameters for each gene as described in the background Section 2.4.4. We use the default threshold of $\alpha = 0.05$ and find that 81% of genes in $\mathbf{D_H}$ and 93% of genes in $\mathbf{D_D}$ are well fitted by the GPD. Upon closer inspection of the worst-fitting estimations, we note that the GPD is still visually close to the true distributions, as seen in Figure A.7. This indicates in some genes there are some peaks in the bins in the tail distributions which the estimation method could not account for, or the spread was too extreme, both reasons hampering the fit. Further attempts to find distinguishing characteristics between the groups of genes that are poorly- and well-fitted in both datasets proved inconclusive. However it is clear that the groups of poorly-fitted genes have a much higher estimated shape parameter than the well-fitted group. Since such high percentages are properly fitted we assume findings can be generalized to the rest of the datasets and proceed with analyses.

3.4.1. Distribution to Network Behaviour Correlations Show the Differences Between HAP1 and DepMap

Now that we have fitted a GPD to each gene, we can more formally correlate behaviours of the distribution to behaviours in the MKNN graphs. We calculate the MKNN graphs for a wide range of k-values, and extract for each node in the graph its degree, its clustering coefficient and its betweenness centrality (defined in Supplementary Table B.1). These connectivity metrics serve as a summary of the graph's behaviour. We then correlate these network behaviours to properties of the well-fitted GPDs. We are interested in large values for k, since nodes have more freedom to show connectivity patterns, but also since a low value for k does not produce enough unique possible data-points to allow for a meaningful correlation of 19385 unique GPD parameters to for example degree or clustering coefficient, which are dependent on k. We therefore performed a grid search over k values from 10 to 200. We show the resulting scatter plots with correlations for both $\mathbf{D_H}$ and $\mathbf{D_D}$ in Supplementary Figures A.10 and A.11 for a representative network $k = 100$. The correlations between the distributions and the degree of nodes over the range of k values is shown in Figure 3.14 and similarly the correlations to the clustering coefficient and betweenness centrality are provided in Supplementary Figures A.8, A.9.

Pertaining to the GPD parameters first. Interestingly the correlations are quite diver-

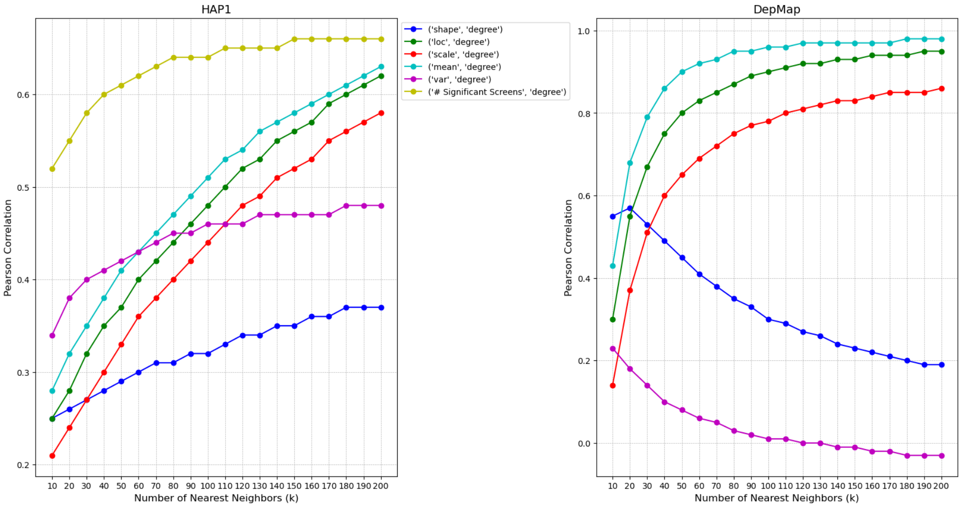


Figure 3.14: Pearson correlations of GLS distribution parameters and the degree of genes in the MKNN networks for k in range [10, 200].

gent between $\mathbf{D_H}$ and $\mathbf{D_D}$. The shape parameter (ξ) has low but increasing correlations with the degree of nodes in $\mathbf{D_H}$, but in $\mathbf{D_D}$ this correlation almost instantly declines as k grows. However, ξ does correlate significantly with the cluster coefficient of nodes in both, especially so in DepMap (see Supplementary Figures A.10 and A.11). Nodes with more heavy-tailed GPDs (higher ξ) are more likely to include extreme events or outliers, and in the networks are correlated to have neighbors which are also more connected amongst each other.

The location parameter (μ or 'loc') essentially translates to the $-\log_{10}$(GLS P value) that a node has for the 90th percentile threshold. Therefore, it is logical for there to be a high correlation between μ and degree. Since a MKNN network is built around reciprocal affinity, when GLS P values are in general higher than they are more likely to be reciprocated between nodes, and nodes with lower values GLS P values are more likely to be disconnected. This is also why the μ and the mean of the distributions are very similar in their correlation values. In DepMap, from $k = 100$ onwards the correlation of 0.95 shows almost exact linear correlation for this trend. But for μ, the cluster coefficient and betweenness centrality correlate opposite to each other in both datasets, where in $\mathbf{D_H}$ nodes with a higher μ tend to be part of well-connected neighborhoods, in $\mathbf{D_D}$ the neighborhoods are almost irrelevant but they are much more important in regulating overall connectivity in the network. The scatter plots (A.10, A.11) also shows that overall the cluster coefficient for nodes in DepMap networks is lower than those in HAP1, while the betweenness centrality is far higher. This shows that while most nodes are connected to the largest component in $\mathbf{D_D}$ (as per the previous section), their neighbours are in general more dissimilar in their interactions strengths. Since betweenness centrality is dependent on the number of nodes in a connected component, it is no surprise that in the disconnected nodes in $\mathbf{D_H}$ lower this value substantially compared to $\mathbf{D_D}$.

Finally for the scale parameter (σ), the higher σ becomes, the greater the spread or dispersion of the distribution. Since a larger dispersion or range can also indicate larger values in the distributions, the correlation behaviour shown by μ and σ are similar in all datasets and correlations. The MKNN networks are more prone to pick the largest connections, and as such, the same assessments for the μ parameter are true for the scale σ.

In the KNN networks, there was a significant difference in behaviour between genes in the core and in the periphery of the networks and found that the number of screens in which genes in those groups score is significantly different (Figure 3.9). We continued the investigation into the effect of the number of scored screens in HAP1, and see in Figure 3.14 and Supplementary Figure A.8 that there is a significant correlation between the degree and to a lesser extent the cluster coefficient of nodes and the number of screens in which a node has a significant P value. Furthermore, we find that over the range of analyzed MKNN networks there are just 6 genes that are enriched for a GO:BP term in the networks that are not significant regulators in any screen.

An investigation to ascertain a correlation between significant screens and network behaviour per gene in DepMap is not trivial, as DepMap uses a slightly different definition for a gene to significantly score. DepMap uses a list of gold-standard nonessential genes from Hart et al. [80] and a list of essential genes which is the intersection of the gold-standard list from Blomen et al. [81] and an essential list again by Hart et al. [80]. It then takes the distribution of CERES scores from the dataset for each gene in both lists to get a distribution for nonessential genes and essential genes. For every other gene not in those lists, it captures the profile of CERES scores and determines a probability for how likely that gene belongs to the either distribution. This raises an important problem in determining when a gene scores significantly. The threshold for this probability would be arbitrary, and no literature presents a suitable one except from a blog post on the DepMap forums where one of the contributors to the Achilles Project suggests 0.5 as a threshold [82]. But this seems rather arbitrary, and due to the different biological contexts in cell lines would most likely have to differ between each screen. A 50% chance of a gene being essential is not robust enough to consider it being significant in a screen, and therefore we do not analyze this metric in the DepMap dataset.

3.4.2. Degree Distributions Capture the Contrast between HAP1 Screens and DepMap

When analyzing the MKNN networks for both datasets over a large range of k values, it became apparent that the network in DepMap consists of a large component which from small k values onwards always contains > 80% of nodes in the dataset, while HAP1 networks tend to have one component of around 1500 nodes and more groups of smaller components, with the bulk of nodes being singular and not interacting with the network. This is clear when analyzing the behaviour of degree distributions of the networks of both datasets over the range of k values. For the MKNN networks they are shown in Supplementary Figures A.12 and A.13. The degree distributions at the midpoint $k = 100$ are shown in Figure 3.15. The distributions are close to mirrored, meaning HAP1 has more genes which tend to have lower degrees while DepMap has more genes with higher degrees. As k increases, the extremes become more and more apparent. The more free-

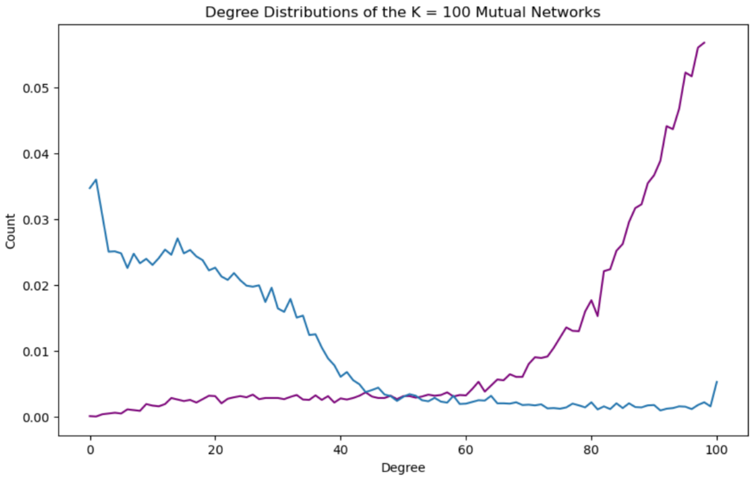


Figure 3.15: Degree distributions for HAP1 (blue) and DepMap (purple) in the MKNN $k = 100$ network.

dom is given, the more nodes in DepMap get a higher degree and the more nodes in HAP1 shift ever so slowly to having a higher degree.

The degree distributions and knowledge of the fragmentation of components provides an explanation for some of the previous observed GPD correlations. The fragmentation into different sized components alters the betweenness centrality of nodes, since fewer nodes are connected by any path. In DepMap the giant component including most nodes with many connections does the opposite, and this is why the correlations in Supplementary Figure A.9 are not significant with any feature of the GPD nor with the number of scored screens in the HAP1 dataset while in DepMap the location and scale grow to correlate almost linearly. As the differences in sizes and connectivity of the components are already present at lower values of k, the correlations around the cluster coefficient are stable in DepMap and increase in HAP1 as the connectivity also increases (Supplementary Figure A.8. We continue investigation by asking *why* these differences in both degree and GPD distributions are present in the datasets.

3.4.3. Control Networks Show Differences in Connectivity Patterns

We asked whether the different network behaviours in terms of connectivity and degree distributions are inherently related to the differences in phenotypes for which the HAP1 screens and the DepMap CRISPR screens are designed for. The trend is that under all network construction methods that DepMap has a much higher percentage of nodes that are connected and there is no distinct separation between a core and periphery, and the degree distributions are mirrored in the datasets (see previous Section). As a control experiment, we selected five groups of genes from separated areas of biology. The groups are genes related to DNA repair, apoptosis, translation, transcription factors and the mitochondria and for each group 100 random genes were selected (selected according to Methodology 2.4.5). The underlying assumption is that it is much more likely for genes within those groups to show interactions in the network, rather than genes interacting between different groups, since this would also be true in a cellular environment.

We construct MKNN networks with these 500 genes over a range of k values from 10 to 100 with steps of 10. Already at $k = 10$, the network is connected in $\mathbf{D_D}$ with all 500 nodes while $\mathbf{D_H}$ remains separated into multiple components, largest of which is 30 nodes. Again, there is fragmentation in $\mathbf{D_H}$ and connection in $\mathbf{D_D}$, and there are already connections between the five distinct groups in both datasets. As a measure of separability of these groups the modularity Q of the network can be calculated. Modularity is also used by the Leiden algorithm, and therefore discussed in the Methodology 2.3.1. Zooming in at an even lower range of k values from range 1 tot 10, the trend in modularity is quickly shown, see Figure 3.16. The division between the five control groups is better than a random subdivision in the HAP1 network at all values of k compared to DepMap, and in addition remains fairly stable. As k increases, and thus the maximum degree nodes can have increases, in DepMap its shown that even these varied groups of biology are become increasingly hard to separate.

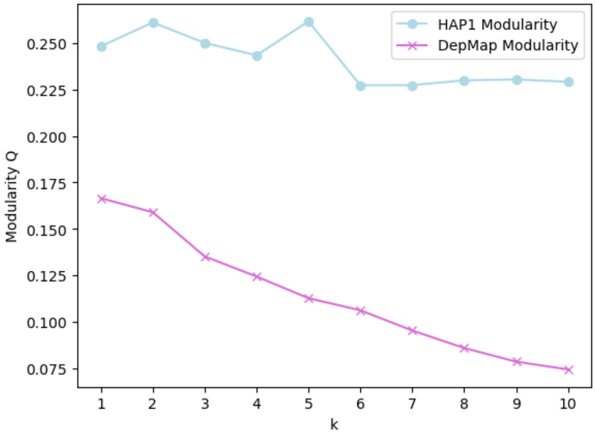


Figure 3.16: Modularity of control networks over a range of MKNN networks.

With the specific protein states measured in the HAP1 screens, it is not expected for most genes to be involved in the regulation or mechanistics of a protein machinery or specific cellular process. In DepMap, synthetic lethality is a much broader phenotype. Already 1500 genes are considered essential by DepMap as defined by Hart and Blomen, and their CERES distributions over all cell lines follow a Gaussian distribution that are highly similar [83]. For such high-level phenotypes it is in general a problem that genes acting in completely different mechanisms can be wrongly associated by the high-level phenotype [13]. The HAP1 screen compendium, even though these five regions of biology were not necessarily screened for, does show how the more specific protein-state readout phenotypes are less likely to associate gene pairs as is the case in DepMap.

3.4.4. Peak at the HAP1 Degree Distributions Implicates Significant Screen Genes

In Figure 3.15, a feature of the degree distribution of the HAP1 dataset is the peak at the maximum degree. This is also pertinent across k values in Figure A.12. The fact that this trend is consistent raised suspicion that this behaviour was due to a data artifact. We analyzed the behaviour of nodes which showed a maximum degree. Notably, seven genes have a maximum degree of k in all networks: DLG2, FHIT, LAMA2, MRPS28, NARS2, PDSS2, TRPM3. Each gene is frequently scored, and is significant by P value in 34 to 117 screens. Furthermore, they are again amongst some of the lengthier genes in the human genome, increasing the odds of being inserted into in the CELL-seq screens, similar to the discussion in the global networks.

From the MKNN $k = 10$ network onwards, we selected a group of 408 genes that were consistently showing a maximum degree in subsequent networks. We compared parameters of the GPD distributions of those 408 genes against the rest of the genes in the $\mathbf{X_H}$ dataset. Results are shown in the violin plot in Figure 3.17. For every tested parameter, shape, scale, mean, variance and number of significant screens, the distribution of values for the groups differed significantly. Genes with more connectivity than most genes thus tend to have distributions in $\mathbf{D_H}$ which are more heavy-tailed, more dispersed, and also generally contain larger interaction strengths.

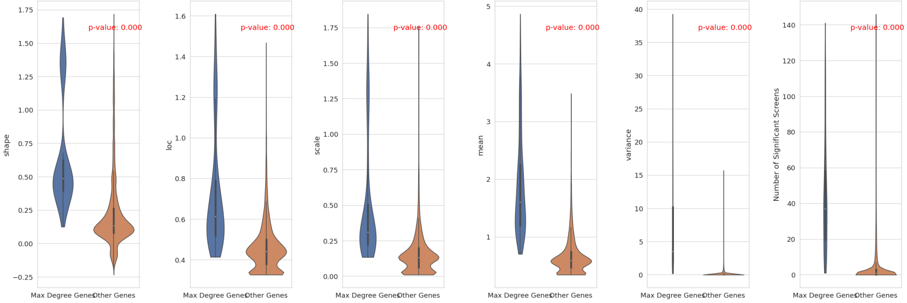


Figure 3.17: Violin plot comparing genes with maximum degree in the network to the remaining genes.

The distribution is most dissimilar when considering the number of screens the genes are scored in. The mean of the maximum degree nodes is almost 40 while for the remaining nodes this is 1 and the distribution is centered around 0. The outliers in the distribution of all genes which show there are genes that score in > 140 screens are those that are connected in the network with a high degree, just never a maximum degree of k. The MKNNs network have thus not removed the hub nodes in its entirety, and they are still showing to be genes that are more frequently significant in the HAP1 screen compendium. Their influence on the network's ability to recapitulate biology has however decreased, as the MKNN networks even with the inclusion of those genes show better recapitulation of biology than the global and KNN networks. The MKNN networks tackled the hub problem, the over-representation of hub nodes, by increasing the fragmentation of the networks and thereby allowing for these hub nodes to be present but still allow meaningful enriched communities to be found. The cost of this fragmentation is that an extremely large groups of nodes remains disconnected from all other nodes. The group of isolated nodes does decrease when increasing k (Supplementary Figure A.15). But networks of higher k values are less biologically relevant in terms of enrichment. The amount of isolated vertices was raised as a problem in both ClusterONE and global network analyses, and we therefore investigate why this is so apparent in MKNN networks as well and attempt a method to resolve this in the following two sections.

3.4.5. Connecting Isolated Nodes Shows Presence of Hub Nodes

To increase connectivity of MKNN graphs, the work by de Sousa, Rezenda and Batista suggest adding an edge between each isolated vertex and its highest ranking nearest neighbor [84], termed the nearest neighbor (NN) approach. This method for increasing connectivity improved results in the work by Dalmia en Sia the least out of additional attempted methods from literature [76], but we argue its interpretability gives insights into why so many nodes remain disconnected.

As an experiment, we applied NN to all isolated vertices over several MKNN networks and gauged whether this improved biological relevance of networks. Additionally we analyzed which nodes these isolated vertices want to connect to the most but are not reciprocated in the MKNN networks. Results are shown in Figure 3.18. At lower more local values of k, it is already apparent how the stringency of the MKNN networks retains a high precision of 0.8 at STRING interactions stronger than 400. Edges present in the network are overrepresented by those who are known to interact, while the additional edges in the NN networks are in lesser numbers represented in STRING, and the additional nodes now participating in the network also lowers the total recall of the network. However, at higher values for k the precision in MKNN networks tends to decrease while in NN networks this actually increases. In other words, as fewer nodes are isolated, it becomes increasingly useful for overall recapitulation of biology to add them to the network. This overall precision-recall curve trend however starts to become more similar to MKNN networks as k becomes larger than 100, as fewer nodes are isolated.

At the $k = 10$ MKNN network with NN we analyzed which new connections were formed. In that network, 55.7% of vertices are isolated. When analyzing to which nodes they are connected using the NN approach, we find that all these isolated vertices want to connect to a similar group of nodes, only 25.6% of the connected network. We can perform a similar analysis as before, and compare the distributions of these nodes' $\mathbf{D_H}$ values to find why these are favoured for connections. We remove the overlapping nodes of maximum degree from the previous section from this set, leaving 22.5% of connected nodes. There is thus little overlap between nodes that are favoured for connection by isolated vertices and those with maximum degree. The results of the comparison are shown in the violin plot in Figure 3.19. The results are similar to the nodes with maximum degree from the previous section. Again, the distributions of the genes which are favoured for connections show heavier tails, higher overall connections, and a larger spread of values. There is again a significant difference between these nodes in terms of the number of screens they score in, albeit less exaggerated than with the maximum degree nodes. The distribution is more centered around those nodes who score in 17 screens.

We conclude from these analyses that there is a subset of nodes which have higher values for all GPD parameters, overall interaction strengths, and are more frequently sig-

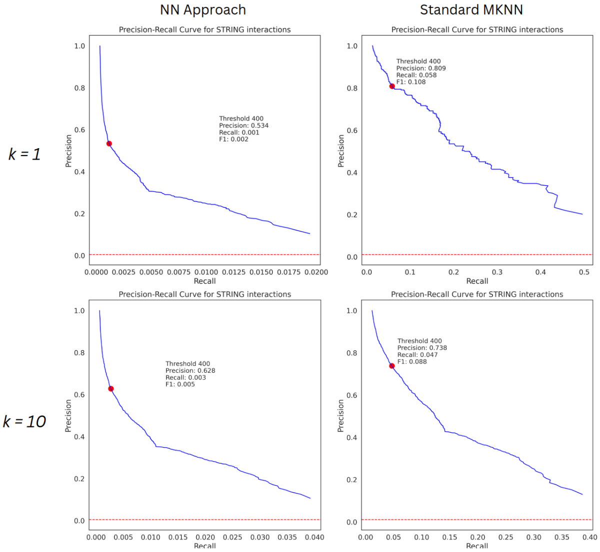


Figure 3.18: Precision-recall curves of STRING interactions in the MKNN networks and in MKNN networks with NN.

nificant regulators in screens which makes them dominant in their influence on network connectivity patterns. It is however difficult to determine if this behaviour is due to underlying biology or from biases in the experimental design. While the presence of hub nodes is a common feature of biological networks generated from most types of biological data, our performed analyses show that their influence in the global and KNN networks are so large, that it is not possible to identify meaningful communities associated to a biological process. We can not with full certainty conclude whether this presence is exaggerated due to experimental design, or whether current approaches to analyze the networks fall short of accommodating for patterns present in actual biology.

We have shown for some genes with hub node behaviour that they are associated with signalling pathways or function in processes where many genes interplay. But genes

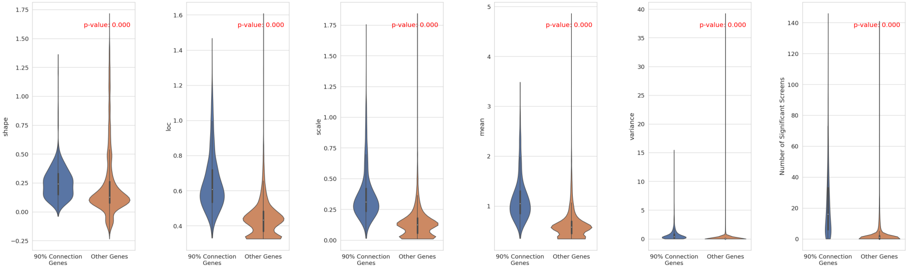


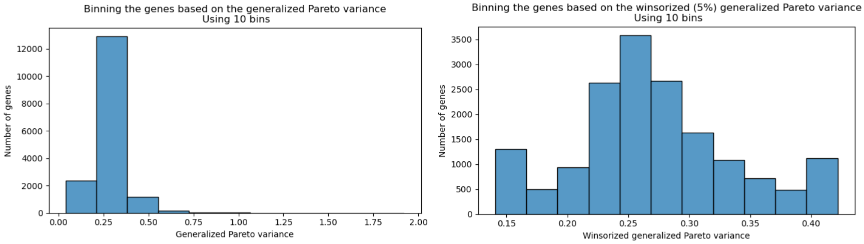
Figure 3.19: Violin plot comparing genes that isolated vertices want to connect to in the network to the remaining genes.

that they are relatively long play are more frequently inserted due to the experimental design. In addition, it is known there are also genes that impact fluorescence levels in the cell by for example regulating cell size are also more frequently found in the high and low populations, thus skewing their representation in the dataset. In the MKNN network's stringency, we limit the degree of hub nodes to at most k with reciprocal affinity and do find more communities and more precisely find interactions also found in curated databases. Even still, the previous analyses on GPD parameters show that scoring is still very important for these nodes to be present in the MKNN networks. In future works, this relation between biological relevance and experimental design should be more intimately investigated in the HAP1 screen compendium.

3.4.6. Distribution Parameter Weighted MKNN Networks

One goal was to incorporate gene specific characteristics to determine a local value for k for each gene. In the cell the amount of interactions per gene product is also not uniform, and the experimental design of the CELL-seq screens has further shown that gene length and their significance in screens is highly associated to network behaviour. We propose a method to take the characteristic of a gene's interaction most associated with its degree, and weigh k based on that characteristic.

We employed the following method to determine a local k: a MKNN network begins with a pre-determined value for k, which will serve as the base value for each node. Then the range of values of the distribution characteristic for all genes are binned into k + 1 bins, labelled from 0 to k. The label of the bin where that node's characteristic lies is added to that node's pre-determined k value. The histogram binning method is sensitive to outliers of the correlated statistic. Therefore the statistic was first winsorized to capture outliers below the 5th and above the 95th percentile. Winsorizing sets all the outliers below or above these percentiles to the value of the percentiles. In practice those values will land in the first or last bin, but the spread between those bins will be more homogeneous since the bin edges are no longer skewed due to outliers as seen in Figure 3.20. The approach ensures that k values become at most twice as much as the base k value for that network, as it has been established networks with lower k values tend to show more enrichment for biological processes.



(a) Naive binning method shows the outliers stretching the x-axis towards extremes making the second bin over-represented.

(b) After winsorizing the statistic, the spread of bins is much more homogeneous and representative of the underlying distribution.

Figure 3.20: Reduction of the effect of outliers leads to more sensible binning used in the weighted-K Nearest Neighbour networks.

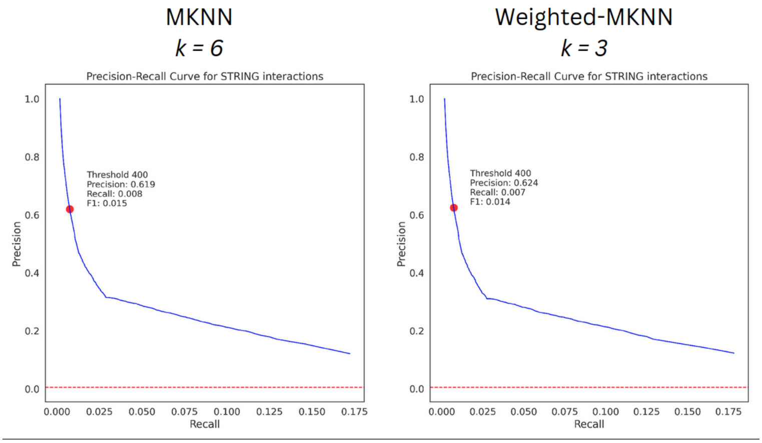


Figure 3.21: STRING precision-recall curves in DepMap Weighted-MKNN and the MKNN.

Observing the graph in Figure 3.14, for the HAP1 dataset the most correlated characteristics with degree are the number of significant screens, the location μ (loc, and associated also the mean), and the scale σ. Though notably, these correlations are not strongly present in the smaller range of k values. Similarly in DepMap, this holds true for μ and σ. We calculate MKNN networks with the above-described k weighing for the range [1, 10]. We show a section of the distributions of GO:BP enrichment P values in Supplementary Figure A.16 for both datasets.

In the HAP1 screen compendium, the resulting networks are not an improvement over a fixed k value for all the most correlated characteristics. The best network was the one weighted by the number of unique screens genes score in. As the effect of weighted k becomes more noticable at higher k values, there is a large drop in the number of communities found to be enriched. This method likely perpetuates the biases towards the number of screens and therefore does not improve. Still, it was the most successful out of all GPD parameters with higher correlations.

In DepMap, we do find a network at weighted-MKNN $k = 3$ with the most amount of communities enriched of all networks thus far when the correlated parameter is μ (see Supplementary Figure A.16) at $k = 3$. This network thus allows networks to have a degree of at most 6. Interestingly, when we compare the STRING scores of this network to that of the MKNN $k = 6$, we find almost identical values in Figure 3.21. So while the overall quality of interactions is very similar, the number of communities enriched does increase, indicating this extra freedom of three k values does appear to improve the networks.

To conclude, the method applied to determine a local k value is a linear method that bins the distribution of the correlated characteristic. Since Pearson correlations quantify the strength and direction of a linear relationship between the variables, it is therefore not surprising that DepMap, who had already better correlations at low k values, shows an improvement where the HAP1 screen compendium does not. Further attempts with a method that assigns a wider range of k did not improve. We continue with a different method to determine a local k value not reliant on the GPD parameters in Section 3.6.

3.5. Analysis on Individual Screen Contributions

An essential consideration in large-scale screening studies is the sufficiency of the number of screens employed, as well as the extent to which these screens contribute independent functional information. We perform a principal component analysis (PCA) on the dataset to generate insights on feature importance. Most interestingly, we find a near-linear relationship between the number of components and the phenotypic variance they explain, shown in Figure 3.22. The first 6 PCs account for 20% of the variance, after which the trend is almost linear for the remaining components, indicating that the dataset contains a relatively consistent pattern of variation across multiple screens, and thus they offer unique information to the dataset. This equal contribution of variation indicates the dataset is limited by the number of screens, since there is no significant decrease in variance explained added by each consecutive PC, meaning the dataset does not lend itself well to dimensionality reduction.

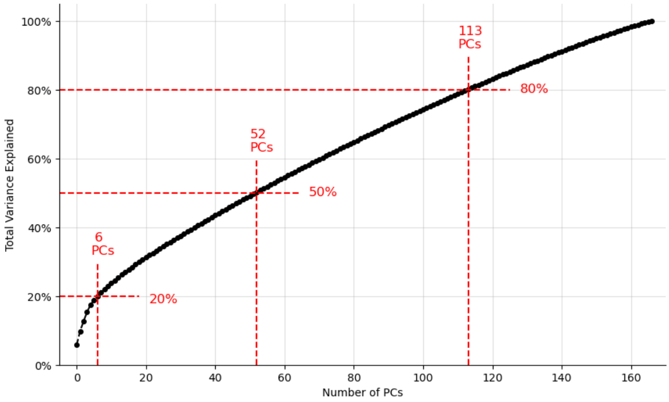


Figure 3.22: The number of Principal Components (PCs) plotted against the amount of variance in the dataset they explain.

In the section on Generalized Least Squares, it was elaborated that the non-uniform coverage of phenotypes lead to related readings in the HAP1 screen compendium which makes the features show nonindependence. To quantify this nonindependence, we investigate the loadings of the components which account for the most variance, and find that the first principal component (PC1) accounts for ~ 6% of the total variance. We can continue to determine the influence of each feature (screen) on the component by obtaining the magnitude of the loading of each feature on PC1, signifying the strength and direction of the relationship. The loadings are shown in Figure 3.23. Phenotypes with the some of the most coverage, RPS6 (10 screens) and lipid transport (9 screens) are indicated by bars with extended lines. As can be clearly seen, their influence on the variance explained by the first component is not only one of the highest, screens within the phenotype are also quite close, indicating screens within the phenotype also contribute similarly.

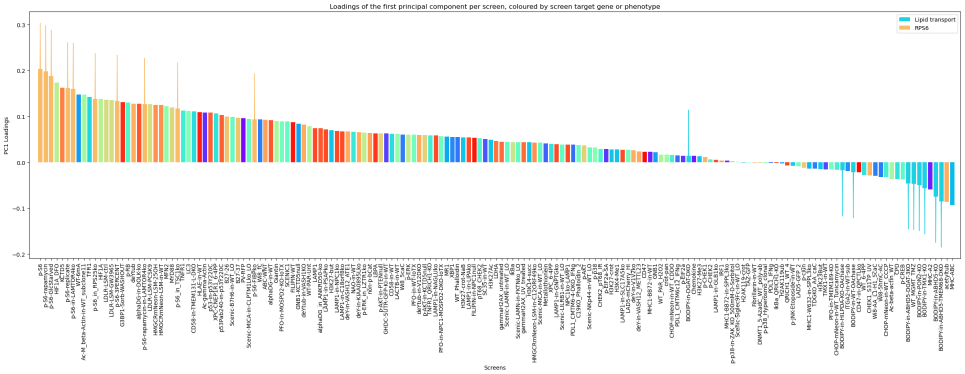


Figure 3.23: The loadings of the first principal component of the HAP1 dataset. Loadings are normalized.

3.5.1. ANALYSIS ON DEPMAP WITH LIMITED SCREENS

We asked whether the difference in network behaviour between the HAP1 screen compendium and DepMap was not merely due to the different biological interpretations, but rather due to the number of screens in DepMap being over 6 times as much as in the HAP1 dataset. To this end, we constructed MKNN networks over a range of k values using splits of the DepMap dataset. We opted to retain coverage of the cancer lineages in the splits since that is one of the strengths of DepMap. Additionally, there are not enough screens from myeloid cancer alone to make a comparison to the CELL-seq screens since the myeloid cell lines would be most similar to HAP1, which is derived from a chronic myeloid leukemia cell line.

As mentioned, DepMap contains 31 distinct cancer lineages, with some lineages only being represented by one or a few cell lines. To retain coverage of cancer types over the splits of screens, we selected screens only for the 15 most present cancer lineages, leaving 879 screens. This amount was used to create five distinct splits of 167 screens, equal to the size of the HAP1 screen compendium. On those five splits we can calculate the GLS distance matrix to obtain five distance matrices $\{D_{D1}, ..., D_{D5}\}$. We track connectivity and functional metrics for the range of MKNN networks for each distance matrix and average the results to deal with potential outlying splits. We opted for five unique splits over a bootstrapping approach with potentially more splits since the computational overhead is infeasible for this project.

The degree distribution of the resulting networks are shown in Figure 3.24. The distributions are the mean distributions over the five splits, but in Supplementary Figure A.14 it is shown the distributions do not vary much over the splits. The degree distributions show the connectivity patterns of the DepMap networks with all screens are conserved when using a reduced number of screens, as the progression over values of k is almost identical to those in DepMap with all screens in Supplementary Figure A.13. The network behaviour is therefore not a result of the greater number of screens that DepMap has when compared to the HAP1 screen compendium.

We next evaluated the quality of the split networks, and report the results of the best split in Figure 3.25. The overall distribution of GO:BP enrichment P values, as well as

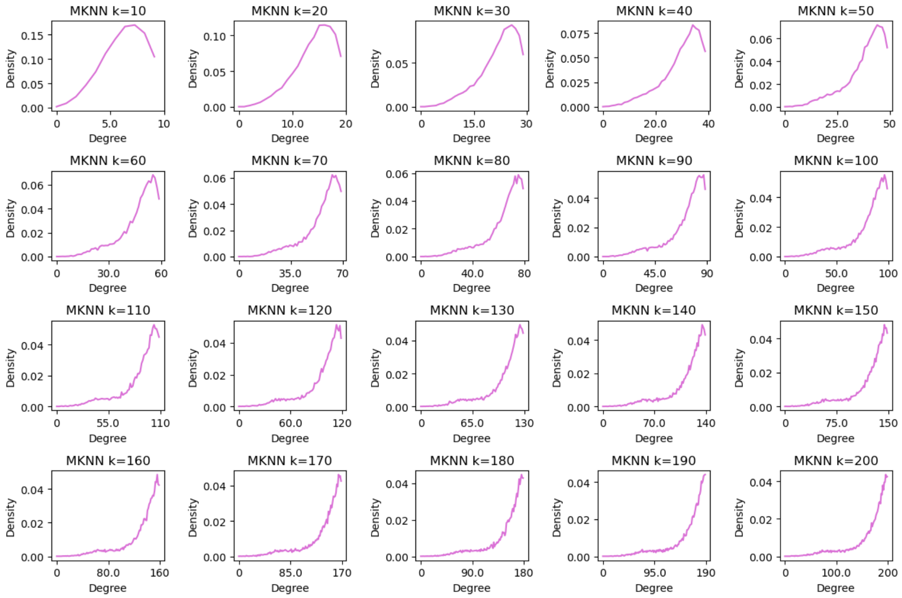


Figure 3.24: Average degree distribution of the networks from five DepMap splits of 167 screens.

the absolute number of enriched communities found is worse across all k values in the best split network when compared to DepMap in Figure 3.12. In addition, the Leiden algorithm was unable to detect communities in the network from $k = 12$ and onwards. This shows how the limited number of screens are less informative and do not allow for networks with any modularity, and the only enriched GO term is the broadest one possible, 'biological process'. While the degree distributions are visually similar to the original DepMap MKNN networks, the networks are not connected such that there are denser subgraphs in the networks. This highlights how the HAP1 screen compendium with its smaller number of screens is able to capture more enriched and more overall communities that reflect biology compared to DepMap with fewer screens.

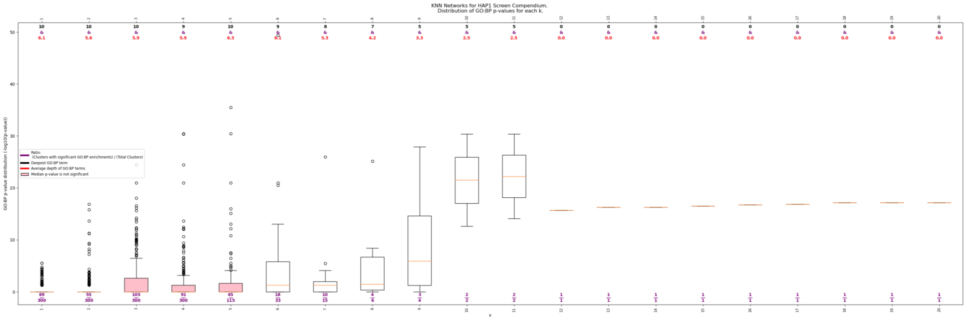


Figure 3.25: Distribution of GO:BP enrichment P values of the best split of 167 DepMap screens.

While community analysis is thus hampered, we can further evaluate the quality of the network using the STRING database (Methodology 2.6.2) and plot the precision-recall curves of known STRING interactions in Figure 3.26. The curves are shown for two k values to highlight the difference over time, as $k = 1$ networks contain fewer edges than $k = 20$ networks this will impact the ability of recall and precision of the networks. The values of precision, recall, and their harmonic mean the F1 score are smaller in the split networks, but the differences are not substantial and their trends are comparable. This indicates that even with a smaller feature size there are still relevant relationships captured in the dataset.

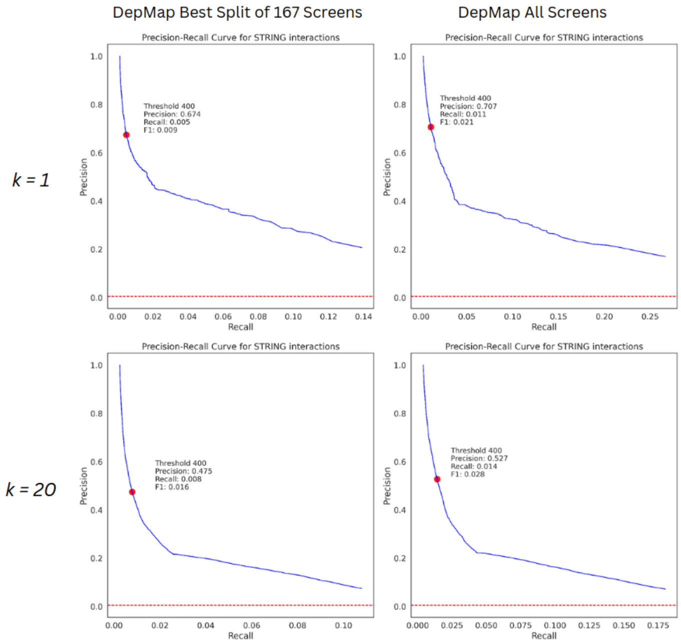


Figure 3.26: Precision recall curves of the DepMap dataset against the split dataset. The highlighted threshold is the point on the PR curve at which STRING interactions with a score of 400 are considered a true interaction.

3.6. Robustness of Neighbour Decrease Shows Cluster Coherence

We next investigated a novel method to construct gene-gene interaction networks with gene-specific thresholds for interactions. Again let $G = (V, E)$ be a graph with every gene being represented as a node $v \in V$. Assume a node v_i is connected to all other nodes found in its distance matrix entry d_i. Similar to a global network, we can increase the threshold for GLS P values along the range found in d_i and note that the number of nodes that v_i interacts with will decrease as the threshold becomes more stringent. We can plot the number of neighbors against the GLS P value threshold for a gene as shown in Figure 3.27. We set an arbitrary number of 1000 steps and count the number of interactions in d_i that cross this threshold. In general, we observe that the drop-off is negligible at first as the threshold is very lenient, then drops very steeply as it not expected for a gene to have an interaction with most other genes, until the drop-off rate diminishes and we are left with a core number of genes still sharing an interaction. We define this section of the diminishing drop-off rate as the most *robust* number of genes that are in a core set of interactions for that gene, as increases in the GLS P value threshold do not greatly influence the amount of genes in this core set. Robustness in this context thus refers to the persistence and reliability of a trend, meaning it has overcome short-term variation.

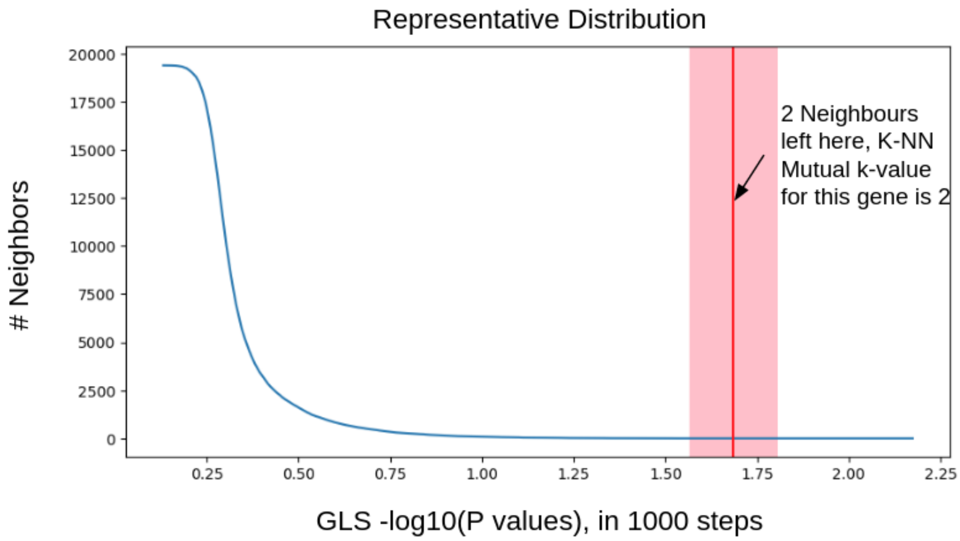


Figure 3.27: Trend of decreasing number of neighbours as the threshold for GLS P values increases. Distribution is representative of most genes from the dataset, differing in the exact range of GLS P values and speed of decrease.

We asked if we can estimate at what threshold the number of neighbors reaches this robust zone for each gene. The number of genes remaining in the core set of a gene's robust are can serve as a value for k in the MKNN networks. Interestingly, this problem translates quite well from different fields. In deep learning, this problem is frequently encountered when determining when a model has converged to stop the training as the validation loss no longer shows significant decrease. In physics, we can interpret the graph in Figure 3.27 as a hill, and when rolling a ball off the side of the hill we can ask when the ball starts to show a stable speed after rolling down the hill. Both of these applications will have a singular point on the axis to determine when the robust zone begins, when using early stopping for the deep learning application [85] or when creating and calculating a free-body diagram for the ball in physics. Since we are more interest in when the neighbor decrease trend is robust, we aim for a solution that accounts for an area of the graph over multiple thresholds points.

For this reason, we apply an approach from economics (and other disciplines) and estimate when the neighbor decrease graph is robust using a moving average (MA). MAs are frequently employed for the smoothing and denoising of time series data as well as trend detection in stock market analysis [86] and has through history been consistently been employed in the development of trading strategies [87]. We define a MA over the GLS P value thresholds similar to the definition by Zakamulin [88], such that the MA uses a fixed-size window that is rolled over the thresholds shown in Equation 3.3.

$$MA_t(W) = \frac{\sum_{t=0}^{W} N_t}{W} \tag{3.3}$$

W is the size of the window, t is the starting threshold and N_t is the number of genes that are above the threshold t for a specific gene, defined as the "neighbors" of a gene at a threshold. There is no optimal size of the window W and it is dataset dependent. To determine when a trend is robust, we use the Moving-average-change-of-direction rule [88], which is defined as $MA_t(W) - MA_{t-1}(W)$. We roll the moving average window across the GLS P value threshold axis and stop the movement when the change-of-direction equals 0, the moment in the change-of-direction rule to change purchasing behaviour. When a MA window is of sufficient size W, we can thus say the core neighborhood has not changed for at least W threshold changes.

To determine a value for W, we use the groups of control genes from Table B.2 for the HAP1 dataset and Table B.3 for DepMap. Per group, we begin with a large window size $W = 250$, a quarter of the maximum window when using 1000 threshold steps. Per gene group, we apply the MA and determine how many genes are left at threshold t when the change-of-direction rule shows 0. If this does not happen for any of the genes in the group, we decrease W by 1 and start again. When all genes have a threshold t for a specific value W, each gene has a core set of neighbors whose GLS P value interaction strength $\geq t$. We define a small undirected network with the genes in the control group as nodes, and add edges between them if a gene is in another gene's core group. Any additional nodes in the core groups of neighbors are also added and connected. If this small network consists of only one component, meaning that we can find a path from each gene in the control group to every other gene in the control group, we have found an appropriate size of W for that control group. Such a network for the first control group is shown in Figure 3.28. In the figure, a window size of 249 connects the control group as there are edges such that there is a path between every gene. Note two additional nodes are added from the core groups.

We perform this analysis for every control group in $\mathbf{D_H}$ and $\mathbf{D_D}$. We find that the largest value for W for a MA such that every control group is connected in $\mathbf{D_H}$ is 54, and for $\mathbf{D_H}$ this is 86. We extrapolate these sizes for W to every gene in the datasets. For each gene we can deduce the threshold t at which the change-of-direction with $MA_t(54)$ or $MA_t(86)$ is 0 for the respective datasets. The number of neighbors in the core group with GLS P value threshold $\geq t$ is the specific k value for that gene in the MKNN network. If the change-of-direction for the MA is never 0, we use a default value for k, such that again a grid search over k can be performed. The resulting distributions of k values are shown in Supplementary Figure A.17.

We again perform a grid search over a range of k from 0 to 10, creating MKNN networks with the k values as shown in Figure A.17 and adding 0 to 10 additional k values in each network to check if the found k values are not too stringent. We define these net-

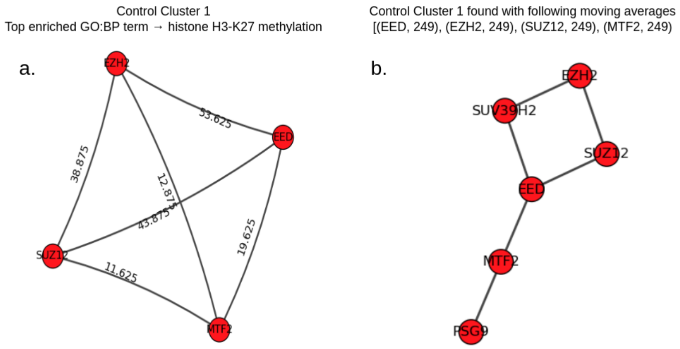

Control Cluster 1
Top enriched GO:BP term → histone H3-K27 methylation
a.
EZH2
EED
SUZ12
MTF2
53.625
38.875
12.875
43.875
19.625
11.625
Control Cluster 1 found with following moving averages
[(EED, 249), (EZH2, 249), (SUZ12, 249), (MTF2, 249)
b.
EZH2
SUV39H2
SUZ12
EED
MTF2
PSG9

Figure 3.28: A: The EED-EZH2-SUZ12-MTF2 control group with $-\log_{10}(P)$ values as interaction strengths. B: The small network found when applying a MA with window size 249. Every gene in the control group is presented as a node along with additional nodes in the core sets. PSG9 is added from the MTF2 core group, SUV39H2 is added from the core groups of both EZH2 and EED.

works as Robust-MKNN networks. The best network for both datasets is shown in Figure 3.29.

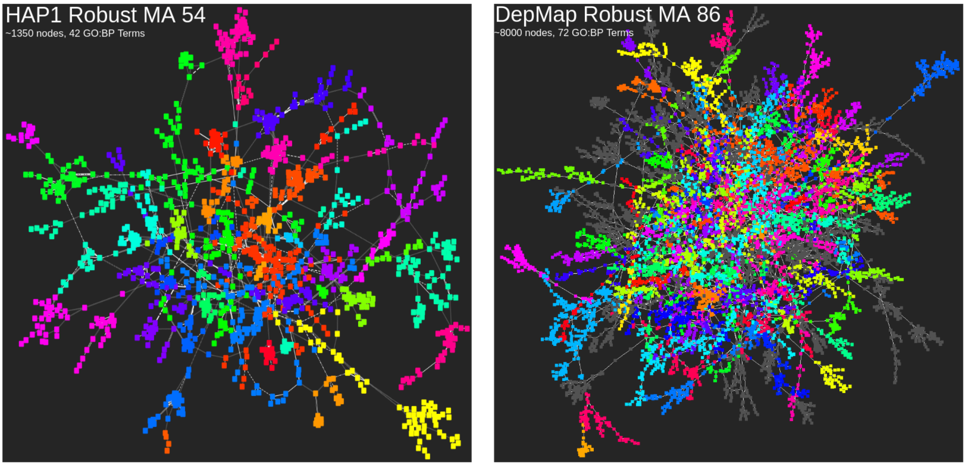


Figure 3.29: Largest component of each dataset's best network based on overall enrichment scores. MA window sizes are 54 and 86 for $\mathbf{D_H}$ and $\mathbf{D_D}$ respectively.

The networks with highest enrichment scores (highest number of GO:BP terms and cluster-agnostic enrichment) are those that do not add additional k values to the nodes in the network. Robust-MKNN does show a different behaviour than previously shown network construction methods. In HAP1, it is overall worse in finding GO:BP enriched communities for all tested k values. But the giant component of Figure 3.29 contains many enriched and separable communities, finding all 12 control clusters while being of larger size than the giant components of regular MKNN networks from k in range [1, 10]. Because the window sizes are defined by the control clusters, finding all twelve is not surprising. If the goal is to create a more connected network whose interactions are more comprehensive of cellular biology, this approach can be promising. However, since its metrics of recapitulating biology are overall lower than standard MKNN networks, as of this moment we do not recommend this approach over MKNN.

In DepMap, the size of the giant component and the number of enriched communities in it is below those found by the MKNN networks as well as in the weighted-MKNN networks, as such this approach is not the best for a desire more connected component network. We do again find all positive controls clustered separately. This worse performance can be due to the the window size not being generalizable to the remaining genes, or the fact that too much freedom is given in terms of k values and we have previously seen the best networks are yielded with lower k values in DepMap. Therefore, as of now this approach is also not recommended for DepMap. Variations on the robust network construction method did not yield better results over the standard MA, therefore recommendations for future work are made in the Conclusion.

3.7. Coverage of Biology

To interpret gene-gene interaction network communities, one of the main methods in bioinformatics is to identify the functional information that best describes the genes in the community. The Methodology (2.3.2) describes how in this work the functional associations of gene groups are associated to terms from the Gene Ontology Biological Processes (GO:BP) using gene set enrichment analysis (GSEA). With this vocabulary of represented GO terms in the networks, we asked what parts of biology are covered in the HAP1 screen compendium, how this compares to the coverage by DepMap, and how the compendium can be expanded to expand its coverage in an optimal manner.

To answer this question, we collected all recapitulated functional categories from the MKNN networks of both the HAP1 screen compendium and DepMap over a range of k values from 1 to 20. For both datasets this amount to a total of 4691 GO:BP terms that communities were functionally enriched for. Such large lists of enriched terms are frequently the results of such analyses, due to the nature of gene annotations in GO. GO, and the knowledge it contains, is a dynamic and ever-expanding compendium of knowledge, and for many gene products not all associations are known or verified. This leads to some genes being annotated to very specific terms lower in the tree structure, while others are only associated to broad terms. GO accounts for this by propagating annotations of genes to all parents of a term so that enrichment analysis can also be successful for genes with non-specific annotations [89].

3.7.1. Naive Higher-Order Comparison Shows Little Differences

To tackle the differences in GO term specificity between terms, we used this propagation of terms property and for every enriched term found the GO terms it is also annotated for at the first and second levels of the hierarchy. Since high level terms cover more general cellular processes, this could give an oversight into coverage of fundamental areas HAP1 screens and DepMap cover. We transformed the list of enriched terms to first and second level terms as described in the background Methodology (2.4.6), and present the results in Supplementary Figure A.18. The two plots show the first two levels of the GO:BP ontological tree, with the inner node being the term "biological process" and the two concentric circles around being the first level (20 terms) and second level (395 terms) of the ontology. The nodes are coloured by the number of times they are found enriched in the networks. The plots both show largest enrichment for the broadest categories in level 1 meaning the categories which capture most terms in the ontology like "cellular process", "biological regulation" and "metabolic process". No differences between the datasets stood out, most terms were in agreement and most number of 2nd level terms showed enrichment in DepMap only.

This comparison did not provide a definite differences between biological coverage of the datasets. As argued by Supek et al., one of the problems of such a visualization is the lack of information these high-level terms provide when performing comparison of GSEA results [90]. Some of the problems include the unbalanced DAG structure, such that some terms have more children and are thus more frequently enriched, lack of information on term specificity differences between datasets, and genes being unevenly annotated which leads to a bias in representation due to some categories simply having more genes annotated [91]. Furthermore, the assymetry of the tree means that for many annotations it is possible to reach many second and first level terms while only ever moving up the tree in different paths, meaning the GO structure is not particularly suited towards finding a singular correct higher-level annotation for a GO term. As such, the visualization in Figure A.18 does not tackle the term specificity problems, and the downstream GO enrichment analysis requires a more sophisticated approach.

3.7.2. Semantic Similarity GO Term Clustering Highlights Enrichment Characteristics

For this reason, this work continued analysis using GO-Compass (Gene Ontology list comparison using Semantic Similarity), a tool for comparing lists of GO terms published in 2023 by Harbig, Paz and Nieselt [89]. The tool is based on semantic similarity, which is a distance measure between GO terms especially developed for the GO structure's assymetry. For this tool, we applied the recommended definition for semantic similarity by the authors of GO-Compass termed the Wang semantic similarity. This measure was introduced in 2007 by Wang et al [92]. The measure defines a distance between two GO:BP terms by considering the set of common ancestors in the ontology alongside their information content, and is therefore much more suited for the comparison of a set of GO:BP terms.

Using this distance, GO-Compass employs hierarchical clustering on the input list of GO terms found from the network analysis, such that groups of highly semantically similar terms are created which are then represented by one parent term, based on the REVIGO algorithm by Supek et al [90]. This hierarchical clustering also weights the found enrichment P values of the found terms. The hierarchical clustering can be manipulated such that the terms at the leaves of the hierarchy are guaranteed to show no value of semantic similarity higher than a pre-determined cut than a cut-off value named *dispensability*. The dispensability of GO terms is defined to be in the range [0, 1], and is anti-correlated with the uniqueness of a GO term in the hierarchical graph, meaning that low dispensability values lead to a term being more unique, "an outlier when compared semantically to the whole list" [90], and vice-versa.

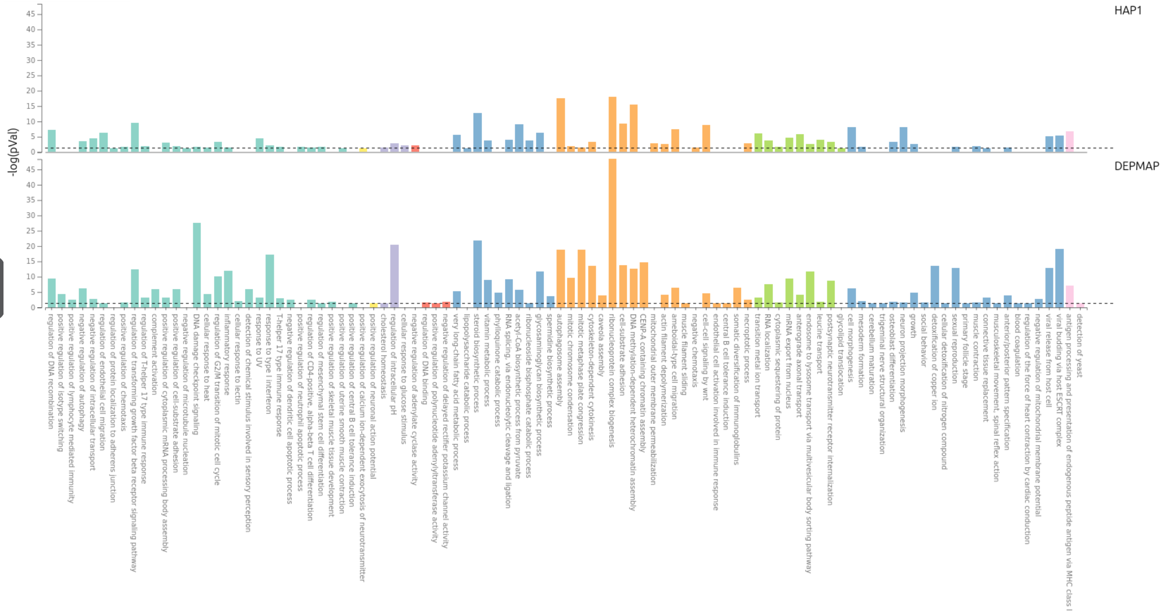


Figure 3.30: Bar graph of enrichment -log10(P values) of 100 GO:BP terms found using GO-Compass.

We set the dispensability cut-off to 0.33 which resulted in a representation of the 4691 unique GO:BP terms to one hundred terms at the leaves of the hierarchy that are guaranteed to be as semantically dissimilar as possible, based on a background ontology and the inputted list of GO:VP terms and their enrichment P values. One hundred terms was chosen as a representative set of GO:BP terms. Other thresholds were examined and future work could change this threshold to include more or fewer terms. 11 of the one hundred terms are unique to the HAP1 dataset, 31 are unique to DepMap and 58 are shared. The bar graph in Figure 3.30 shows the enrichment values (in log-scale) of the captured GO:BP terms from the networks. The bar graph is coloured by 10 different groups which indicate each group shares a common ancestor GO-term in the hierarchical clustering not present in other groups. These ancestors are the 10 GO:BP terms most semantically dissimilar when using a dispensability cut-off of 0.1 on the hierarchical clustering tree, and function to separate the 100 terms into 10 distinct areas of biology. The dotted-line represent the threshold for significance ($\alpha < 0.05$) which is approximately 1.3 in log-scale.

Overlap Between Datasets

The overlap is quite remarkable, 27 of the top 30 most enriched terms in DepMap are also captured in the HAP1 networks, with "CENP-A containing chromatin assembly", "detoxification of copper-ion" and "somatic diversification of immunoglobulins" not present in HAP1. It is not surprising for these terms to be more present in DepMap than in HAP1. CENP-A is a histone variant which provides structural integrity to the chromatin in cells and is present in the centromere region of a chromosome. Its function is important for chromosome segregation during cell division [93]. As the HAP1 cells are haploid, this separation of chromosomes is not a part of mitosis, the process of cell division, and therefore the CENP-A reliant chromatin assembly is logically less represented in HAP1 cells. Detoxification of copper-ion is a a GO term related to an important regulatory function of cellular reaction to copper ion, which is a micronutrient essential to the function of cells and human health [94]. Its essentiality indicates it should score high in the DepMap dataset in particular. Finally, the somatic diversification of immunoglobulins is an essential function of cells, since this GO term is related to the ability of antibodies to be modified to specific antigens. Antibodies are exclusively produced by B cells. DepMap covers a wide range of cancer cell lines, and some of these cell lines are B cells or related to B cell lymphomas[2]. Vice-versa, the top 30 most enriched GO:BP terms in HAP1 are all enriched in DepMap.

The graph shows for most groups that enrichment P values in DepMap are generally higher, indicating DepMap is better at representing those GO terms in the networks. However, statistical overrepresentation is also dependent on the size of the background gene set, and since DepMap contains about 2000 fewer genes than the HAP1 screen compendium, enrichment values will always be not perfectly comparable. It is however promising that with a limited number of screens the overlap in higher-level GO terms between the datasets is so large.

GO Terms Unique to the HAP1 Screen Compendium

Eleven GO:BP terms found in the networks are unique to the HAP1 screen compendium, and are highlighted in Supplementary Figure A.19. One of the goals of this work is to bring forward hypotheses for biological interactions not present in literature or other datasets. Therefore, these eleven unique GO:BP terms are a promising set of interactions to investigate further. In addition, the networks' proven capacity to generate biologically meaningful clusters opens up possibilities for attributing gene functions to genes present within the these found communities, even if not all genes or their interactions with the community documented in the literature [4]. Therefore, we identified the associated networks and communities belonging to each unique enriched term, and investigated which genes of the community are annotated to that term and whether these interactions are sensible from a biological perspective and whether they can present novel interactions. Note that all genes found in the 11 GO:BP communities are present in both datasets and the uniqueness is therefore not derived from a lack of presence in DepMap.

For seven of the found GO terms, the enriched GO terms do not lend themselves for the discovery of interactions unique to biology or to the HAP1 screen compendium for that matter. Two GO terms were associated with communities of > 30 genes, but only two or three genes in those communities were annotated for that GO term. The GO terms for these communities thus show barely passing significance for enrichment just below the significance threshold of $\alpha = 0.05$, and GO terms at a greater depth (more specific) are much more suited to describe these groups, but those are also found in DepMap. Another two enrichment's which are not too meaningful are related to pairs of genes, which are both only found in $k = 1$ networks. As these networks grow, these pairs of genes join larger communities which are highly enriched for a specific GO term shared

[2] https://depmap.org/portal/cell_line/ACH-000070?tab=overview

between both datasets. These pairs are thus only found as the most closest related gene pairs, and are enriched for a more broader GO term at barely significant threshold simply as a consequence of the strict $k = 1$ network. However, they are also recapitulated in the STRING database as known interactions. Another three unique GO:BP terms stem from a similar issue, where a pair or triplets of genes are annotated to a higher level GO:BP term not present in DepMap, to which only one or two genes are annotated in the database, but a more specific GO term for which all two or three are annotated exists that is shared in DepMap and also a better representative. This could indicate that some of those genes should contain an annotation in the GO term database to the higher-level term, which could result from the varying annotations bias in GO as mentioned earlier. All of these interactions are known in literature so their annotation to the same GO:BP term could be sensible, however it is not fruitful for potential new interactions found in the HAP1 screen compendium.

However, the remaining four unique GO terms come from singular gene-gene pairs that are present in networks with relatively high k values > 5, for which there is either little evidence in known literature or in DepMap but are sensible in biological context. They are listed in Table 3.3. Note that each pair is enriched with a barely significant P value, a more stringent cut-off for enrichment would not have found these GO terms. However their conserved presence as pairs over a range of k in the MKNN networks, robust reciprocal affinity, and potential hypotheses for interactions warrant further investigation.

GO:BP Term	P value	Enriched Community	GO Annotated Gene	$-log_{10}$ (GLS P Value)
glycolipid translocation	0.0467	RFT1, DNAJC11	RFT1	0.879
lipopolysaccharide catabolic process	0.0496	AOAH, PPP2R2B	AOAH	5.837
negative regulation of microtubule nucleation	0.0496	ARHGEF7, PAK2	ARHGEF7	1.420
positive regulation of uterine smooth muscle contraction	0.0490	GPALPP1, GPER1	GPER1	0.202

Table 3.3: Enriched interactions unique in the HAP1 screen compendium.

Glycolipid translocation. RFT1 is a gene coding for the RFT1 protein whose function is related to the transport of oligosaccharide (multiple carbohydrates) through the membrane of the endoplasmic reticulum (ER), a cellular subunit [95]. DNAJC11 encodes for a protein involved in the organization of the mitochondrial membrane and it is especially involved with the Mitochondrial Contact Site and Cristae Organizing System (MICOS) complex [96]. The ER and mitochondria have been shown to interact in the cell for various purposes, and their main form of interaction is due to contact between their membrane structures [97]. Contact sites on their membranes are termed mitochondria-associated ER membranes (MAMs). One of the primary involvements of MAMs are lipid transfer between the two organelles and calcium signalling where calcium is transported from the calcium-rich ER to the mitochondria. Oligosaccharides can facilitate in the protein-protein interactions required for the MAM to perform its functions. The MICOS complex is involved with the architecture of the mitochondrial membrane including formation of the MAM and thus the physical tethering of the ER to the mitochondria. Therefore, the interaction of RFT1 and DNJAC11, both in their location and function seems logically supported by cellular biology. However, the interaction strength is not that high and neither is the enrichment P value, furthermore they are not involved in the STRING database. The pair's biological context, its uniqueness to HAP1 and the fact that these genes only interact with each other up until $k = 5$ networks do warrant closer examination in a laboratory setting.

Lipopolysaccharide catabolic process. The HAP1 dataset shows the highest interaction strength of the four uniquely enriched gene pairs between AOAH and PPP2R2B. Their interaction is not found in STRING or CORUM, and using g:profiler they are not enriched in any other major database. However they have been mentioned together in the work by Xiong et al. which studies the expression profiles of genes in B cells during the progression of Alzheimer's Disease (AD), wherein PPP2R2B and AOAH are both in the top 18 of upregulated genes as AD progresses [98]. They identify PPP2R2B as known to be involved with AD, but AOAH's inclusion is not further explored. AOAH is a lipide, an enzyme related to fats, and most closely associated to detoxify lipopolysaccharides (LPS) on the outer membrane of gram-negative bacteria, a sub-class of bacteria [99]. Interestingly, both LPS on its own and gram-negative bacteria are involved in several pathologies related to AD, and they are being suggested as targets for therapeutics to treat AD [100]. AOAH role in LPS detoxification is critical for the immune's response to bacterial infections, and its role in AD as well as PPP2R2B's known role in neural function make this interaction a strong candidate for further analysis in a laboratory setting.

Negative regulation of microtubule nucleation. The interaction between ARHGEF7 and PAK2 shows the highest possible STRING score of 999. Their involvement with the cytoskeletal organization was already hypothesized in the beginning of the 21st century and has been later experimentally confirmed. The GO term for which they are enriched is quite specific to microtubule creation for which ARHGEF7 is annotated, and while PAK2 is involved with the overall organization of the cytoskeleton of which microtubules are an important part, no literature supported PAK2 in their creation. Since their interaction is well-studied, the GO term is most likely accurate and its exclusion of PAK2 warranted.

Positive regulation of uterine smooth muscle contraction. The interaction strength between GPALPP1 and GPER1 is quite low, however they are connected to only each other until $k = 10$, after which they join into a larger community of 41 nodes that is enriched but these two genes are not associated to that GO term. Their low interaction strength do not indicate that the screens overall show these genes as a meaningful interaction. However, it is interesting that this interaction is not documented in literature but these two genes both play a role in AD, similar to the second relation in Table 3.3. The literature on GPALPP1 is limited, however it is mentioned in a recent June 2023 study by Wu et al. as a biomarker for the diagnosis of AD [101]. GPER1 on the other hand is a well studied endrogen receptor protein that is associated to interact with female sex hormones to regulate menstrual cycles. For this reason it is associated to the unique GO:BP term. A 2019 review by Roque et al. does implicate GPER1 as being associated with several diseases among which is AD, citing experimental trials in mice models and some literature suggesting it being used as a therapeutic target to treat AD [102]. The biological context therefore makes this interaction not entirely unlikely, both being associated to AD, but more evidence would be required to substantiate their interaction. Coupled with the low GLS P value score this interaction is worth mentioning but should not be the first interaction to investigate experimentally.

3.7.3. Expansion of the HAP1 Screen Compendium

One of the goals of this work was to bring forward recommendations on how to expand the HAP1 screen compendium in an optimal fashion. This section will elaborate on the difficulties of selecting an optimal expansion and lists some approaches. However, we can make direct recommendations for expansion which flow from the found interactions unique to the HAP1 screen compendium brought forward in the previous section. The most promising additional screens are those that investigate the RFT1-DNAJC11 and the AOAH-PPP2R2B interactions. It would be interesting to perform a differential analysis for those four proteins. Four screens can be performed which screen for those proteins. Then, similarly to other screens in the compendium, HAP1 cells that have one of those genes knocked-out can be cloned for an additional four knockout screens. This would allow direct comparison between the activity of genes that are regulators for those phenotypes when the gene is present and when it is not.

Besides direct knockout analyses, additional screens could also be performed for the processes related to these interactions. The RFT1-DNJAC11 interaction is most likely present in the compendium due to nine screens being related to lipid transport. A more targeted phenotype to this interaction would be more appropriate as to not further contribute to the nonindependence of screens. Further research could be done into proteins involved in the MAM, which is vital to understand the interactions between the mitochondria and endoplasmic reticulum and this is further supported by the work of Xia et al. who note the limited study in organelle-organelle interaction and the importance of the interaction of the mitochondria with other organelles in cancer [97]. As mentioned, one of the functions of the MAM is calcium signalling, which is the transport of calcium ions for cell signalling involved in a multitude of processes. Two potential screens could be for ITPR3 and SIGMAR1, which are both proteins located at the MAM interface and path of the calcium signalling pathways. For the AOAH-PPP2R2B interactions, a screen for a phenotype that could include their involvement is more difficult, as they already have a highly significant interaction in the GLS P value matrix. The literature implicates them being present during development of AD and specifically points to the involvement of gram-negative bacteria in AD, but this is a multi-faceted area of many protein-protein interactions contributing and selecting one or two phenotypes is not feasible without intimate knowledge of involved processes. Of course, the experimental validation is not limited to CELL-seq screens and could be explored using other techniques to measure biological interactions like gene expression or mass spectrometry.

Instead of recommendations for screens brought forward by the data, an approach for expansion could be to screen for phenotypes that make the compendium more similar to other well-studied datasets such as DepMap. This could in aid in future comparisons of their relative performance in network analysis. Using DepMap, it could be interesting to add screens for proteins that are involved in the GO:BP terms that are found in DepMap but not in HAP1. However, for many of these terms it would be quite difficult to do so. The three examples of terms in DepMap and not in HAP1 that are most enriched are there due to the nature of the cell lines in DepMap and the essentiality phenotype. Certain processes simply cannot be recovered due to only using one cell line, where cellular activity is particular to that cell type. In addition, processes that could be shown in HAP1 which are enriched only in DepMap are mostly those related to cell survival. DepMap gives insights into these processes since that is its experimental design. In CELL-seq screens, it is very difficult to screen for phenotypes essential to survival. Insertions into essential genes are hard to measure since an insertion leads to the essential gene's protein not being expressed, and this could lead to cell death. The insertions can only be counted and sorted with FACS on live cells.

Another method would be to expand coverage of the GO:BP tree as much as possible. Since the GO is an expansive representation of cellular biology, hitting most categories using the screens is a method to track progress in covering all cellular processes. A starting point could be three broader level 1 terms of the GO:BP tree not covered by the compendium; "biological phase", "biological process involved in intraspecies interaction between organisms" and "detoxification". Biological phase and detoxification are related to process with repetitive processes such as mitosis or immune response and removal of toxic molecules, both mostly associated with essential processes which are again hard to screen for. Intraspecies interaction is a term related to social behaviours, applicable to organisms which show group behaviour. About 60 genes are annotated to this term in humans and screens could thus potentially target this category. The problem is with ticking of categories in the ontology is that associations are very unbalanced, and even at the same level terms can be associated with many or a few genes. The expansiveness of the tree and the relations between terms that jump levels of hierarchy and the multiple paths between children and parents make it hard to identify optimal categories for coverage. The literature does not propose a solution as of yet for this problem.

Finally, an important consideration for potential supplementary screenings is the observed divergence in behavior between genes that demonstrate significance and exhibit a $\log_2(MI) \geq |1|$, compared to those that do not meet these criteria. Notably, genes that do not score in the compendium display a markedly reduced rate of being integrated into networks and enriched communities. This distinction is exemplified by the subset of 8090 genes that are common to both HAP1 and DepMap datasets but do not score in any screening. Prioritizing the design of phenotypes that are under the regulation of genes within this list of 8090 shared genes could alleviate this divergence. In the process of deliberating what phenotypes to screen for, an initial phase would involve formulating hypotheses about genes with higher likelihoods of scoring in the screen based on knowledge of the phenotype. An additional step could entail cross-referencing these postulated genes with the aforementioned compilation of 8090 genes. This cross-referencing step serves as an adjunctive metric for gauging the potential contribution of conducting the proposed screen.

4

Conclusion and Discussion

4.1. Conclusion

This study is the culmination of a scientific exploration aimed at assessing the capabilities of the compendium of haploid phenotype screens to be integrated for the purpose of creating a gene-gene interaction network. The construction of biological networks is integral to broadening understanding of cellular interactions and ultimately hopes to identify targets for treatment of diseases with cellular therapies. Such networks are primarily based on biological datasets such as gene expression, pairwise gene knockout screens, protein-protein interaction databases or genome-wide interaction profiles associated to high-level phenotypes such as essentiality or growth. The unique haploid screens designed for measuring varied cellular phenotypes have not before been applied in such contexts. Therefore this work is exploratory in nature. The guiding line in this work aims to establish whether known techniques are directly applicable to the dataset and systematically exploring interpretable and novel approaches suited to the characteristics of the dataset.

The efforts towards this goal are centered around four research questions, with at the center of the inquiry:

Can we leverage the dataset to create a gene-gene interaction network that shows proof of relevancy in current literature on genetic interactions?

To answer this main question, we applied two approaches from literature designed for clustering and gene-gene interaction network inference to serve as a benchmark, ClusterONE and ARACNE-AP. ClusterONE showed that the HAP1 dataset does contain smaller densely-connected groups of genes that recapitulate known biology, while ARACNE-AP's information theoretic approach found a network without structural characteristics, making it unsuited for biological interpretation. Since the dataset has unique properties not presently evaluated by other approaches in literature, we next sought to investigate the dataset's network capabilities more systematically by applying known and interpretable approaches for network creation. As such, we created networks that only include interactions above a pre-defined threshold, and showed that very strong interactions strengths are limited to a subset of nodes and that this was a requirement for networks to be informative. In a KNN network approach, only at the most immediate interaction of $k = 1$ we find dense clusters associated to biology, but as k increased to 2 and above the network lost most structural properties that allow for enrichment of groups of genes as well as an overabundance of interactions not found in databases of curated real interactions. We conclude the dataset contains a core-periphery structure where a smaller group of highly inter-connected genes is juxtapositioned by a large subset genes that exhibit of sparse connectivity.

As such, we moved on to an approach that maintains the focus on local connectivity, while being used in literature to target such core-periphery structures: mutual KNN (MKNN). The application of MKNN on the dataset captured the most number of associated biological processes and had high precision in capturing interactions that are known to reflect real biological interactions. The structure of the MKNN networks positively impacted the dense core structure by limiting the number of interactions, which led to networks with few nodes that interact and also many isolated vertices.

We continued the analysis into why the networks presented these behaviours by fitting a generalized Pareto distribution on each genes' distribution on the distance matrices, and find that nodes that show higher connectivity in the networks are those with a heavy-tailed and more dispersed distribution, along with higher average interaction strengths. Importantly, we also found that there is a discrepancy in behaviour between genes that are scored as a regulator in many phenotype screens versus those that never or rarely score. While some of the genes that score the most are those associated to many biological interactions, the experimental design is also biased towards finding those genes that are of longer length or influence fluorescence levels when the cell populations are sorted. As the networks that scored the best are those that limited the influence of those genes with the most connectivity, we hypothesize that the behaviour of those genes is not merely due to biological reasons but goes in conjunction with biases brought forward by the experimental design.

The aforementioned approaches are based on a threshold or k value that is equally applied to all nodes in the network. However, one value does not capture the intricate range of dynamics present in a cellular environment, and therefore two approaches were newly presented to predict an optimal value for k in the MKNN networks. Unfortunately, the GPD parameter weighted MKNN approach worsened the networks in the HAP1 screen compendium, and arguably did not improve the MKNN networks from DepMap a considerable amount. The linear relationship between the GPD parameters and the networks did not hold true in the HAP1 dataset for small values of k. The second approach, termed robust-MKNN, did furthermore not improve upon a standard MKNN structure. However, its application to the HAP1 dataset did result in a network with the largest number of interactions in a singular component that showed separability of the component into several densely structured groups annotated to different biological processes. As such, this shows potential for a future fine-tuning of this method to generate a more comprehensive cellular wiring map.

One of the goals of this work was to identify novel gene-gene interactions that are unique to the CELL-seq screens brought forward by the gene-gene interaction networks.

We have shown the HAP1 screen compendium contains genes with similar phenotypic profiles that in combination are able to recapitulate known biological processes in constructed networks. Furthermore, interactions in the networks show that those with the strongest interactions are more likely to also score high in the STRING PPI database as well as in the CORUM protein complex database. The networks' proven capacity to generate biologically meaningful clusters opens up possibilities for attributing gene functions to genes present within the networks, even if their interactions are not documented in the literature. As such, we identified two gene pairs that i) associate over a wide range of MKNN networks, ii) whose interactions are unique to the HAP1 screens and not present in curated databases, and iii) do show a potential interaction when examining their biological context. The interaction between the RFT1 and DNAJC11 gene products is hypothesized to be related to the contact site between the membranes of the mitochondria and the endoplasmic reticulum. Another interaction conserved over networks in the CELL-seq screens is the relationship between AOAH and PPP2R2B. They have recently been discovered in literature as possible bio-markers for the progression of Alzheimer's disease (AD), and their functional annotations additionally place them in processes related to AD. We therefore recommend designing an experiment to more thoroughly investigate these specific interactions, or experimentally verifying them through additional screens.

How does the haploid phenotype screen compendium compare to DepMap, whose dataset is antagonistic to ours by screening one phenotype over more than a thousand cell lines?

The network analyses showed varying behaviours in the HAP1 screen compendium and DepMap networks, as well as differences in the biological processes they recover. Most notably, we find that in the most suited approach for HAP1 analysis, mutual KNN networks, the behaviours in terms of connectivity are close to mirrored. The highest proportion of genes in DepMap is connected with a high degree, while in HAP1 the opposite of true. The networks in DepMap show they are quickly inclined to create one giant component, while the HAP1 network consists of fragmented subgraphs. The conducted experiments show that the behaviour of the networks is dictated by the experimental design of the screens, where DepMap screens for a high-level phenotype and HAP1 screens for diverse cellular phenotypes.

We furthermore investigated the differences in biological coverage between the datasets as defined by their coverage of the gene ontology, a database which represents the structure of biological processes in cells. We find that to a large degree the retrieved processes are in concordance, while the overall enrichment shows that DepMap has a higher significance for the shared processes. We do find processes that are uniquely associated to the HAP1 and DepMap datasets. These unique processes were further investigated and eventually led to the proposal of potentially new interactions described above.

Does the limited compendium screen size fit the data-hungry task of cellular mapping?

A limitation of the dataset is the number of screens being small, which reduces its statistical power and hampers the accuracy of found interactions. A PCA analysis of the dataset shows that there is an almost linear trend between the number of screens and the percentage of variance they explain. Most screens contribute equally to the variance and almost all screens are required to explain 90% of variance, which shows the importance of each screens inclusion in the dataset and how impactful the reduction of screens would be. We advocate for the expansion of the dataset in order to increase confidence in the shown interactions.

We further investigated if the large discrepancy in the number of screens, 167 for HAP1 and 1078 for DepMap, was an additional reason for the differences in network behaviours. The conducted experiment split the DepMap data into five groups of 167 screens that retain the coverage of diverse cancer lineages, and again construct MKNN networks using these splits and average their results. This shows that networks created with all 167 HAP1 screens are better at recapitulating biological processes and interactions compared to DepMap when DepMap is limited to 167 screens. The above-mentioned structural behaviours remained intact, but the limited number of screens did hugely impact the ability to recall known biology. The HAP1 compendium derived networks are therefore competitive despite the limited sample size.

How can we optimally add screens to most effectively increase diversity in the phenotypic space?

The limited number of screens in the compendium narrows its ability to capture the full range of biological contexts and additional screens could thus benefit the ability of networks to show crucial interactions relevant in more contexts than those currently screened for. In addition, new screens can also more closely validate novel interactions that are brought forward by the current compendium. One of the aims of this expansion of CELL-seq screens is to further investigate potential interactions that are unique to the dataset. For these reasons we recommended additional screens to be performed that investigate the gene products of the genes involved in brought-forward possible novel interactions, RFT1-DNAJC11 and AOAH-PPP2R2B. It is important to investigate further if these interactions are brought forward by the nonindependence, and if screening for these interactions further exaggerate this nonindependence. A secondary approach would be to screen for phenotypes not directly concerning those interactions but rather the underlying processes, which are mitochondrial-endoplasmic reticulum pathways and processes associated to AD. If these interactions are not suited for investigation, and rather the expansion of phenotypic diversity and mediation of nonindependence is the goal, this work does not conclusively establish an approach. There are problems related to increasing diversity based on the gene ontology structure, as well as problems in screening for phenotypes to make the compendium's coverage more akin to that of DepMap. However, we do recommend that when hypothesizing potential targets for future screens, it is important to weigh those screens with more importance if they are more likely to score previously unscored genes.

4.2. LIMITATIONS

This work has attempted to answer the research questions in an exploratory manner using a network-based approach. In comparing to DepMap, we limited the comparisons to behavioural or functional patterns of the networks that were brought forward by the approaches employed in this work. DepMap however is a well-established dataset and other works create networks or perform clustering approaches that are more tailored towards DepMap and are therefore much more successful in creating networks. For example, when showing HAP1 is competitive to DepMap when DepMap has a limited number of screens, this could have been a feature of the MKNN approach not being optimal for DepMap instead of a strength of the HAP1 screen compendium. Furthermore, we are limited in our analyses by only viewing the HAP1 screens through the network perspective.

Another limitation of this work is that it is built on some assumptions which are not fully verified. The effect of the preprocessing approaches on the dataset are not thoroughly examined here. Since the networks do show relevant biology, we assume correctness of these approaches but do not quantify to which extent they aid or limit the networks. This could have helped in reasoning about whether the observed network behaviour stem from biology, experimental design or preprocessing methods. Furthermore, in the Robust-MKNN networks we assumed that the found window to ascertain robustness could be found on the control clusters and then generalized to the remaining genes. However, the analysis on the GPD's have shown that distributions between strong connectors and other genes are quite varied, and a more suitable approach to determining robustness for all genes could have benefited the Robust-MKNN networks.

In addition, one primary method of assessing network quality and biological coverage was based on the functional enrichment of found communities in the network. However, this method is highly dependent on the clustering algorithm. While network quality was also assessed in cluster-agnostic manners, biological coverage was fully dependent on this. The Leiden algorithm used in this work was chosen due to its interpretability and presence in literature. It is also highly dependent on a resolution parameter that can be adapted to, in general, recuperate communities of larger or smaller sizes. We acknowledge that for some networks that show less modularity or fragmentation a grid-search should have been performed for this resolution parameter. This grid-search was performed for some MKNN networks but not systematically, and a more thorough investigation would have certainly aided to the conclusiveness of found results. Finally, the expansion of the screen compendium by this cluster-based approach with coverage did not yield conclusive approaches to expand the phenotype screens to cover most biology. A more thorough comparison of coverage in other known databases or with other datasets could have mitigated that.

4.3. Future Research Recommendations

One of the primary findings of this work is the relatedness between a genes behaviour in networks and the number of times it scored as a regulator in the phenotype screens. This work does implicate the experimental design of the CELL-seq screens towards this behaviour, and shows that when corrected for, the overall network quality improves. However, it is important to investigate whether the hub node behaviour of this genes is entirely due to the measurement technique, a reflection of biology, or both. While we hypothesize experimental design as the culprit, a more thorough investigation into this could aim to potentially correct for the experimental biases towards those over represented genes. This could be in the form of a preprocessing step prior to the applications of a network construction method. When it is in detail known what the biases are and how they can best be dealt with, one research avenue is to explore whether the CELL-seq screen compendium can be complimentary to other datasets like DepMap. Perhaps integration of the CELL-seq screens with other datasets can aid in expand the coverage of biology.

Initially, one of the aims of this work was to also include causal inference on the direction of interactions in the networks or include hierarchical organization of biology. For the inference of hierarchy, we did attempt to include analyses using a tool developed by the Ideker lab for this purposed titled DDOT [103]. However, the application of this algorithm was unfeasible due to its long running time and memory requirements. Even an application of a smaller set of genes solely from the mitochondria did not produce any results. We propose a more thorough investigation into why this popular algorithm is not applicable to the HAP1 screens. For further causal inference, the application of ARACNE-AP showed the difficulties of estimating a probability distribution with the limited number of screens for the calculation of mutual information. Since mutual information is commonly used in causal inference strategies, we recommend to investigate whether the features of the distribution of GLS P values would instead more suited for estimating mutual information. The memory requirements were too large for all genes when calculating them for ARACNE-AP, but could be feasible for a subset of genes.

In the approach for estimating when a decrease in neighbors was robust in Section 3.6, we used a simple moving average (MA). However, recent literature from trend estimation in finance [87], [88] as well as denoising tasks of frequently employed deep learning optimizers such as ADAM employ more sophisticated approaches to the MA, such as exponentially weighted MA or a double-crossover MA. In addition the change-of-direction indicator can be too strict in some cases and rather a percentage drop off could be employed like in most early stopping approaches. We've experimented with both additional mentioned MAs, as well as more lenient change-of-direction indicators, but never found a better alternative than the simple MA employed in the work. It is of course highly dependent on the control groups used, and the assumption that a MA with the right parameters (type, window size, change rule) can be extrapolated from those fitted controls to the entire dataset. Because the approach did show meaningful clustering, we suggest a more precise deep-dive into this method of finding a local threshold for genes in this application.

Bibliography

[1] S. Mukherjee, "Song of the cell: An exploration of medicine and the new human," in Scribner, 2023, ch. We Shall Always Return to the Cell.

[2] A.-L. Barabasi and Z. N. Oltvai, "Network biology: Understanding the cell's functional organization," *Nature reviews genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[3] K. Luck, D.-K. Kim, L. Lambourne, *et al.*, "A reference map of the human binary protein interactome," *Nature*, vol. 580, no. 7803, pp. 402–408, 2020.

[4] G. Turco, C. Chang, R. Y. Wang, *et al.*, "Global analysis of the yeast knockout phenome," *Science Advances*, vol. 9, no. 21, eadg5702, 2023. DOI: 10.1126/sciadv.adg5702. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.adg5702. [Online]. Available: https://www.science.org/doi/abs/10.1126/sciadv.adg5702.

[5] M. K. Yu, M. Kramer, J. Dutkowski, *et al.*, "Translation of genotype to phenotype by a hierarchy of cell subsystems," *Cell systems*, vol. 2, no. 2, pp. 77–88, 2016.

[6] T. A. Manolio, F. S. Collins, N. J. Cox, *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[7] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An expanded view of complex traits: From polygenic to omnigenic," *Cell*, vol. 169, no. 7, pp. 1177–1186, 2017.

[8] R. Liu, C. A. Mancuso, A. Yannakopoulos, K. A. Johnson, and A. Krishnan, "Supervised learning is an accurate method for network-based gene classification," *Bioinformatics*, vol. 36, no. 11, pp. 3457–3465, Apr. 2020, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa150. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/11/3457/50670844/bioinformatics_36_11_3457.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaa150.

[9] M. Costanzo, A. Baryshnikova, J. Bellay, *et al.*, "The genetic landscape of a cell," *Science*, vol. 327, no. 5964, pp. 425–431, 2010. DOI: 10.1126/science.1180823. eprint: https://www.science.org/doi/pdf/10.1126/science.1180823. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1180823.

[10] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[11] B. Rauscher, F. Heigwer, L. Henkel, T. Hielscher, O. Voloshanenko, and M. Boutros, "Toward an integrated map of genetic interactions in cancer cells," *Molecular Systems Biology*, vol. 14, no. 2, e7656, 2018. DOI: https://doi.org/10.15252/msb.20177656. eprint: https://www.embopress.org/doi/pdf/10.15252/msb.20177656. [Online]. Available: https://www.embopress.org/doi/abs/10.15252/msb.20177656.

[12] M. Brockmann, V. A. Blomen, J. Nieuwenhuis, *et al.*, "Genetic wiring maps of single-cell protein states reveal an off-switch for gpcr signalling," *Nature*, vol. 546, no. 7657, pp. 307–311, 2017.

[13] J. M. Replogle, R. A. Saunders, A. N. Pogson, *et al.*, "Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq," *Cell*, vol. 185, no. 14, 2559–2575.e28, 2022, ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2022.05.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867422005979.

[14] Y. Wang and N. E. Navin, "Advances and applications of single-cell sequencing technologies," *Molecular cell*, vol. 58, no. 4, pp. 598–609, 2015.

[15] J. E. Carette, C. P. Guimaraes, M. Varadarajan, *et al.*, "Haploid genetic screens in human cells identify host factors used by pathogens," *Science*, vol. 326, no. 5957, pp. 1231–1235, 2009. DOI: 10.1126/science.1178955. eprint: https://www.science.org/doi/pdf/10.1126/science.1178955. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1178955.

[16] T. Bürckstümmer, C. Banning, P. Hainzl, *et al.*, "A reversible gene trap collection empowers haploid genetics in human cells," *Nature methods*, vol. 10, no. 10, pp. 965–971, 2013.

[17] R. A. Weinberg and R. A. Weinberg, "The biology of cancer," in WW Norton & Company, 2006, ch. The Biology and Genetics of Cells and Organisms.

[18] J. E. Carette, M. Raaben, A. C. Wong, *et al.*, "Ebola virus entry requires the cholesterol transporter niemann–pick c1," *Nature*, vol. 477, no. 7364, pp. 340–343, 2011.

[19] G. Giaever and C. Nislow, "The yeast deletion collection: A decade of functional genomics," *Genetics*, vol. 197, no. 2, pp. 451–465, 2014.

[20] N. J. O'Neil, M. L. Bailey, and P. Hieter, "Synthetic lethality and cancer," *Nature Reviews Genetics*, vol. 18, no. 10, pp. 613–623, 2017.

[21] P. Essletzbichler, T. Konopka, F. Santoro, *et al.*, "Megabase-scale deletion using crispr/cas9 to generate a fully haploid human cell line," *Genome research*, vol. 24, no. 12, pp. 2059–2065, 2014.

[22] M. Cvijovic, J. Almquist, J. Hagmar, *et al.*, "Bridging the gaps in systems biology," *Molecular genetics and genomics : MGG*, vol. 289, Apr. 2014. DOI: 10.1007/s00438-014-0843-3.

[23] O. D. Iancu, A. Colville, P. Darakjian, and R. Hitzemann, "Chapter four - coexpression and cosplicing network approaches for the study of mammalian brain transcriptomes," in *Brain Transcriptome*, ser. International Review of Neurobiology, R. Hitzemann and S. Mcweeney, Eds., vol. 116, Academic Press, 2014, pp. 73–93. DOI: https://doi.org/10.1016/B978-0-12-801105-8.00004-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128011058000047.

[24] A. Tsherniak, F. Vazquez, P. G. Montgomery, *et al.*, "Defining a cancer dependency map," *Cell*, vol. 170, no. 3, 564–576.e16, 2017, ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2017.06.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867417306517.

[25] W. Bonner, H. Hulett, R. Sweet, and L. Herzenberg, "Fluorescence activated cell sorting," *Review of Scientific Instruments*, vol. 43, no. 3, pp. 404–409, 1972.

[26] R. M. Meyers, J. G. Bryan, J. M. McFarland, *et al.*, "Computational correction of copy number effect improves specificity of crispr–cas9 essentiality screens in cancer cells," *Nature genetics*, vol. 49, no. 12, pp. 1779–1784, 2017.

[27] R. Kundra, H. Zhang, R. Sheridan, *et al.*, "Oncotree: A cancer classification system for precision oncology," *JCO Clinical Cancer Informatics*, no. 5, pp. 221–230, 2021, PMID: 33625877. DOI: 10.1200/CCI.20.00108. eprint: https://doi.org/10.1200/CCI.20.00108. [Online]. Available: https://doi.org/10.1200/CCI.20.00108.

[28] M. Wainberg, R. A. Kamber, A. Balsubramani, *et al.*, "A genome-wide atlas of co-essential modules assigns function to uncharacterized genes," *Nature genetics*, vol. 53, no. 5, pp. 638–649, 2021.

[29] K. Shimada, J. A. Bachman, J. L. Muhlich, and T. J. Mitchison, "Shinydepmap, a tool to identify targetable cancer genes and their functional connections from cancer dependency map data," *Elife*, vol. 10, e57116, 2021.

[30] O. H. Iwenofu, R. D. Lackman, A. P. Staddon, D. G. Goodwin, H. M. Haupt, and J. S. Brooks, "Phospho-s6 ribosomal protein: A potential new predictive sarcoma marker for targeted mtor therapy," *Modern pathology*, vol. 21, no. 3, pp. 231–237, 2008.

[31] A. Aitkin, "On least squares and linear combination of observations," *Proceedings of the Royal Society of Edinburgh*, vol. 55, pp. 42–48, 1935.

[32] S. Tenny and M. R. Hoffman, *Odds Ratio*. StatPearls Publishing, Treasure Island (FL), 2022. [Online]. Available: http://europepmc.org/books/NBK431098.

[33] M. Gillespie, B. Jassal, R. Stephan, *et al.*, "The reactome pathway knowledgebase 2022," *Nucleic Acids Research*, vol. 50, no. D1, pp. D687–D692, Nov. 2021, ISSN: 0305-1048. DOI: 10.1093/nar/gkab1028. eprint: https://academic.oup.com/nar/article-pdf/50/D1/D687/42058295/gkab1028.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkab1028.

[34] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, 1996, pp. 226–231.

[35] R. Benfeitas, *Introduction to biological network analysis.*

[36] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, "Gene co-expression analysis for functional classification and gene–disease predictions," *Briefings in Bioinformatics*, vol. 19, no. 4, pp. 575–592, Jan. 2017, ISSN: 1477-4054. DOI: 10.1093/bib/bbw139. eprint: https://academic.oup.com/bib/article-pdf/19/4/575/25193126/bbw139.pdf. [Online]. Available: https://doi.org/10.1093/bib/bbw139.

[37] V. A. Traag, L. Waltman, and N. J. Van Eck, "From louvain to leiden: Guaranteeing well-connected communities," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[38] U. Brandes, D. Delling, M. Gaertler, *et al.*, "On modularity clustering," *IEEE transactions on knowledge and data engineering*, vol. 20, no. 2, pp. 172–188, 2007.

[39] L. Heumos, A. C. Schaar, C. Lance, *et al.*, "Best practices for single-cell analysis across modalities," *Nature Reviews Genetics*, pp. 1–23, 2023.

[40] A. Baryshnikova, "Systematic functional annotation and visualization of biological networks," *Cell Systems*, vol. 2, no. 6, pp. 412–421, 2016, ISSN: 2405-4712. DOI: https://doi.org/10.1016/j.cels.2016.04.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S240547121630148X.

[41] M. Ashburner, C. A. Ball, J. A. Blake, *et al.*, "Gene ontology: Tool for the unification of biology. the gene ontology consortium," en, *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000.

[42] A. Baryshnikova, M. Costanzo, Y. Kim, *et al.*, "Quantitative analysis of fitness and genetic interactions in yeast on a genome scale," *Nature methods*, vol. 7, no. 12, pp. 1017–1024, 2010.

[43] F. Emmert-Streib, G. V. Glazko, G. Altay, and R. de Matos Simoes, "Statistical inference and reverse engineering of gene regulatory networks from observational expression data," en, *Front. Genet.*, vol. 3, p. 8, Feb. 2012.

[44] U. Raudvere, L. Kolberg, I. Kuzmin, *et al.*, "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)," *Nucleic Acids Research*, vol. 47, no. W1, W191–W198, May 2019, ISSN: 0305-1048. DOI: 10.1093/nar/gkz369. eprint: https://academic.oup.com/nar/article-pdf/47/W1/W191/28879887/gkz369.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkz369.

[45] J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo, "G: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments," *Nucleic acids research*, vol. 35, no. suppl_2, W193–W200, 2007.

[46] D. Szklarczyk, A. L. Gable, D. Lyon, *et al.*, "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, vol. 47, no. D1, pp. D607–D613, Nov. 2018, ISSN: 0305-1048. DOI: 10.1093/nar/gky1131. eprint: https://academic.oup.com/nar/article-pdf/47/D1/D607/27437323/gky1131.pdf. [Online]. Available: https://doi.org/10.1093/nar/gky1131.

[47] D. Szklarczyk, A. L. Gable, K. C. Nastou, *et al.*, "The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic Acids Research*, vol. 49, no. D1, pp. D605–D612, Nov. 2020, ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1074. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D605/40395991/gkaa1074.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkaa1074.

[48] J. Pan, R. M. Meyers, B. C. Michel, *et al.*, "Interrogation of mammalian protein complex structure, function, and membership using genome-scale fitness screens," *Cell systems*, vol. 6, no. 5, pp. 555–568, 2018.

[49] G. Tsitsiridis, R. Steinkamp, M. Giurgiu, *et al.*, "CORUM: the comprehensive resource of mammalian protein complexes–2022," *Nucleic Acids Research*, vol. 51, no. D1, pp. D539–D545, Nov. 2022, ISSN: 0305-1048. DOI: 10.1093/nar/gkac1015. eprint: https://academic.oup.com/nar/article-pdf/51/D1/D539/48440639/gkac1015.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkac1015.

[50] E. A. Boyle, J. K. Pritchard, and W. J. Greenleaf, "High-resolution mapping of cancer cell networks using co-functional interactions," *Molecular Systems Biology*, vol. 14, no. 12, e8594, 2018. DOI: https://doi.org/10.15252/msb.20188594. eprint: https://www.embopress.org/doi/pdf/10.15252/msb.20188594. [Online]. Available: https://www.embopress.org/doi/abs/10.15252/msb.20188594.

[51] B. Albert-László and P. Márton, *Network Science*. Cambridge University Press, 2017.

[52] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 2006. [Online]. Available: https://igraph.org.

[53] Gustavsen, J. A., Pai, *et al.*, "Rcy3: Network biology using cytoscape from within r," *F1000Research*, 2019. DOI: 10.12688/f1000research.20887.3.

[54] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks," *Nature methods*, vol. 9, no. 5, pp. 471–472, 2012.

[55] A. Lachmann, F. M. Giorgi, G. Lopez, and A. Califano, "ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information," *Bioinformatics*, vol. 32, no. 14, pp. 2233–2235, Apr. 2016, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw216. eprint: https://academic.oup.com/bioinformatics/article-pdf/32/14/2233/19568312/btw216.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btw216.

[56] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, “The mutual information: Detecting and evaluating dependencies between variables,” *Bioinformatics*, vol. 18, no. suppl_2, S231–S240, 2002.

[57] H. Zhang, S. Kiranyaz, and M. Gabbouj, “Data clustering based on community structure in mutual k-nearest neighbor graph,” in *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, IEEE, 2018, pp. 1–7.

[58] S. Rath, R. Sharma, R. Gupta, *et al.*, “Mitocarta3. 0: An updated mitochondrial proteome now with sub-organelle localization and pathway annotations,” *Nucleic acids research*, vol. 49, no. D1, pp. D1541–D1547, 2021.

[59] S. A. Lambert, A. Jolma, L. F. Campitelli, *et al.*, “The human transcription factors,” *Cell*, vol. 172, no. 4, pp. 650–665, 2018.

[60] A. Subramanian, P. Tamayo, V. K. Mootha, *et al.*, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005. DOI: 10.1073/pnas.0506580102. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.0506580102. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.0506580102.

[61] D. Klopfenstein, L. Zhang, B. S. Pedersen, *et al.*, “Goatools: A python library for gene ontology analyses,” *Scientific reports*, vol. 8, no. 1, p. 10 872, 2018.

[62] B. M. Kuenzi and T. Ideker, “A census of pathway maps in cancer systems biology,” *Nature Reviews Cancer*, vol. 20, no. 4, pp. 233–246, 2020.

[63] A. A. Margolin, I. Nemenman, K. Basso, *et al.*, “Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context,” in *BMC bioinformatics*, BioMed Central, vol. 7, 2006, pp. 1–15.

[64] K. Drew, C. Lee, R. L. Huizar, *et al.*, “Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes,” *Molecular Systems Biology*, vol. 13, no. 6, p. 932, 2017. DOI: https://doi.org/10.15252/msb.20167490. eprint: https://www.embopress.org/doi/pdf/10.15252/msb.20167490. [Online]. Available: https://www.embopress.org/doi/abs/10.15252/msb.20167490.

[65] E. Almaas, A.-L. Barabasi, E. Koonin, Y. Wolf, and G. Karev, “Power laws in biological networks,” in Jun. 2007, pp. 1–11, ISBN: 978-0-387-25883-6. DOI: 10.1007/0-387-33916-7_1.

[66] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, *et al.*, “Using graph theory to analyze biological networks,” *BioData mining*, vol. 4, no. 1, pp. 1–27, 2011.

[67] D. R. Zerbino, P. Achuthan, W. Akanni, *et al.*, “Ensembl 2018,” *Nucleic acids research*, vol. 46, no. D1, pp. D754–D761, 2018.

[68] I. Lopes, G. Altab, P. Raina, and J. P. De Magalhães, “Gene size matters: An analysis of gene length in the human genome,” *Frontiers in Genetics*, vol. 12, p. 559 998, 2021.

[69] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures," *Social networks*, vol. 21, no. 4, pp. 375–395, 2000.

[70] R. Miao and T. Li, "Informative core identification in complex networks," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 85, no. 1, pp. 108–126, Jan. 2023, ISSN: 1369-7412. DOI: 10.1093/jrsssb/qkac009. eprint: https://academic.oup.com/jrsssb/article-pdf/85/1/108/49347906/qkac009.pdf. [Online]. Available: https://doi.org/10.1093/jrsssb/qkac009.

[71] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018. DOI: 10.21105/joss.00861. [Online]. Available: https://doi.org/10.21105/joss.00861.

[72] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, no. 5569, pp. 910–913, 2002.

[73] P. Csermely, A. London, L.-Y. Wu, and B. Uzzi, "Structure and dynamics of core/periphery networks," *Journal of Complex Networks*, vol. 1, no. 2, pp. 93–123, Oct. 2013, ISSN: 2051-1310. DOI: 10.1093/comnet/cnt016. eprint: https://academic.oup.com/comnet/article-pdf/1/2/93/7093753/cnt016.pdf. [Online]. Available: https://doi.org/10.1093/comnet/cnt016.

[74] Z. Hu and R. Bhatnagar, "Clustering algorithm based on mutual k-nearest neighbor relationships," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 5, no. 2, pp. 100–113, 2012.

[75] D. Sardana and R. Bhatnagar, "Graph algorithm to find core periphery structures using mutual k-nearest neighbors," *International Journal of Artificial Intelligence and Applications (IJAIA)*, vol. 12, no. 1, 2021.

[76] A. Dalmia and S. Sia, "Clustering with UMAP: why and how connectivity matters," *CoRR*, vol. abs/2108.05525, 2021. arXiv: 2108.05525. [Online]. Available: https://arxiv.org/abs/2108.05525.

[77] K. Ozaki, M. Shimbo, M. Komachi, and Y. Matsumoto, "Using the mutual k-nearest neighbor graphs for semi-supervised classification on natural language data," in *Proceedings of the fifteenth conference on computational natural language learning*, 2011, pp. 154–162.

[78] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[79] B. Thijssen and L. F. A. Wessels, "Approximating multivariate posterior distribution functions from monte carlo samples for sequential bayesian inference," *PLOS ONE*, vol. 15, no. 3, pp. 1–25, Mar. 2020. DOI: 10.1371/journal.pone.0230101. [Online]. Available: https://doi.org/10.1371/journal.pone.0230101.

[80] T. Hart, K. R. Brown, F. Sircoulomb, R. Rottapel, and J. Moffat, “Measuring error rates in genomic perturbation screens: Gold standards for human functional genomics,” *Molecular Systems Biology*, vol. 10, no. 7, p. 733, 2014. DOI: https://doi.org/10.15252/msb.20145216. eprint: https://www.embopress.org/doi/pdf/10.15252/msb.20145216. [Online]. Available: https://www.embopress.org/doi/abs/10.15252/msb.20145216.

[81] V. A. Blomen, P. Májek, L. T. Jae, *et al.*, “Gene essentiality and synthetic lethality in haploid human cells,” *Science*, vol. 350, no. 6264, pp. 1092–1096, 2015. DOI: 10.1126/science.aac7557. eprint: https://www.science.org/doi/pdf/10.1126/science.aac7557. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aac7557.

[82] J. Dempster, *Depmap genetic dependencies faq*, Jul. 2020. [Online]. Available: https://forum.depmap.org/t/depmap-genetic-dependencies-faq/131.

[83] J. M. Dempster, I. Boyle, F. Vazquez, *et al.*, “Chronos: A cell population dynamics model of crispr experiments that improves inference of gene fitness effects,” *Genome biology*, vol. 22, pp. 1–23, 2021.

[84] C. A. R. de Sousa, S. O. Rezende, and G. E. Batista, “Influence of graph construction on semi-supervised learning,” in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, Springer, 2013, pp. 160–175.

[85] L. Prechelt, G. Montavon, G. Orr, and K.-R. Muller, “Early stopping - but when?” In *Neural networks: Tricks of the Trade Second Edition*. Springer Berlin Heidelberg, 2012, ISBN: 9783642352898.

[86] A. Raudys, “Optimal negative weight moving average for stock price series smoothing,” in *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, 2014, pp. 239–246. DOI: 10.1109/CIFEr.2014.6924079.

[87] C.-H. Park and S. H. Irwin, “What do we know about the profitability of technical analysis?” *Journal of Economic Surveys*, vol. 21, no. 4, pp. 786–826, 2007. DOI: https://doi.org/10.1111/j.1467-6419.2007.00519.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6419.2007.00519.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6419.2007.00519.x.

[88] V. Zakamulin, “Market timing with a robust moving average,” *SSRN Electronic Journal*, May 2015. DOI: 10.2139/ssrn.2612307.

[89] T. Harbig, M. W. Paz, and K. Nieselt, “Go-compass: Visual navigation of multiple lists of go terms,” *Computer Graphics Forum*, vol. 42, no. 3, pp. 271–281, 2023. DOI: https://doi.org/10.1111/cgf.14829. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14829. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14829.

[90] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, “Revigo summarizes and visualizes long lists of gene ontology terms,” *PloS one*, vol. 6, no. 7, e21800, 2011.

[91] C. C. Sosa, D. C. Clavijo-Buriticá, V. H. García-Merchán, *et al.*, "Gocompare: An r package to compare functional enrichment analysis between two species," *Genomics*, vol. 115, no. 1, p. 110 528, 2023, ISSN: 0888-7543. DOI: https://doi.org/10.1016/j.ygeno.2022.110528. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888754322002737.

[92] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, Mar. 2007, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm087. eprint: https://academic.oup.com/bioinformatics/article-pdf/23/10/1274/49812526/bioinformatics_23_10_1274.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btm087.

[93] C. Renaud-Pageot, J.-P. Quivy, M. Lochhead, and G. Almouzni, "Cenp-a regulation and cancer," *Frontiers in Cell and Developmental Biology*, vol. 10, 2022, ISSN: 2296-634X. DOI: 10.3389/fcell.2022.907120. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcell.2022.907120.

[94] I. Scheiber, R. Dringen, and J. F. Mercer, "Copper: Effects of deficiency and overload," *Interrelations between essential metal ions and human diseases*, pp. 359–387, 2013.

[95] M. A. Haeuptle, F. M. Pujol, C. Neupert, *et al.*, "Human rft1 deficiency leads to a disorder of n-linked glycosylation," *The American Journal of Human Genetics*, vol. 82, no. 3, pp. 600–606, 2008.

[96] F. Ioakeimidis, C. Ott, V. Kozjak-Pavlovic, *et al.*, "A splicing mutation in the novel mitochondrial protein dnajc11 causes motor neuron pathology associated with cristae disorganization, and lymphoid abnormalities in mice," *PloS one*, vol. 9, no. 8, e104237, 2014, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0104237. [Online]. Available: https://europepmc.org/articles/PMC4128653.

[97] M. Xia, Y. Zhang, K. Jin, Z. Lu, Z. Zeng, and W. Xiong, "Communication between mitochondria and other organelles: A brand-new perspective on mitochondria in cancer," *Cell & Bioscience*, vol. 9, no. 1, pp. 1–19, 2019.

[98] L.-L. Xiong, L.-L. Xue, R.-L. Du, *et al.*, "Single-cell rna sequencing reveals b cell–related molecular biomarkers for alzheimer's disease," *Experimental & molecular medicine*, vol. 53, no. 12, pp. 1888–1901, 2021.

[99] R. Singh, Y.-L. Chen, S. W. Ng, *et al.*, "Phospholipase activity of acyloxyacyl hydrolase induces il-22-producing cd1a-autoreactive t cells in individuals with psoriasis," *European Journal of Immunology*, vol. 52, no. 3, pp. 511–524, 2022. DOI: https://doi.org/10.1002/eji.202149485. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eji.202149485. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/eji.202149485.

[100] S. Kim, S. J. Shin, Y. H. Park, *et al.*, "Gram-negative bacteria and their lipopolysaccharides in alzheimer's disease: Pathologic roles and therapeutic implications," *Translational Neurodegeneration*, vol. 10, no. 1, pp. 1–23, 2021.

[101] W. Wu, G. Chen, Z. Zhang, M. He, H. Li, and F. Yan, "Construction and verification of atopic dermatitis diagnostic model based on pyroptosis related biological markers using machine learning methods," *BMC Medical Genomics*, vol. 16, no. 1, pp. 1–15, 2023.

[102] C. Roque, J. Mendes-Oliveira, C. Duarte-Chendo, and G. Baltazar, "The role of g protein-coupled estrogen receptor 1 on neurological disorders," *Frontiers in Neuroendocrinology*, vol. 55, p. 100 786, 2019, ISSN: 0091-3022. DOI: https://doi.org/10.1016/j.yfrne.2019.100786. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0091302219300482.

[103] M. K. Yu, J. Ma, K. Ono, *et al.*, "Ddot: A swiss army knife for investigating data-driven biological ontologies," *Cell systems*, vol. 8, no. 3, pp. 267–273, 2019.

A

Supplementary Figures

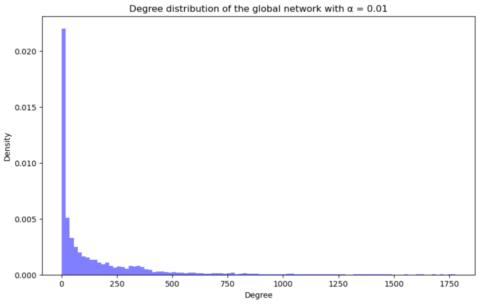


Figure A.1: Degree distribution of the f_{global} network using $\alpha = 0.01$. Density is for a degree d is calculated as $\frac{n_d}{n}$, the fraction of nodes with degree d over all nodes.

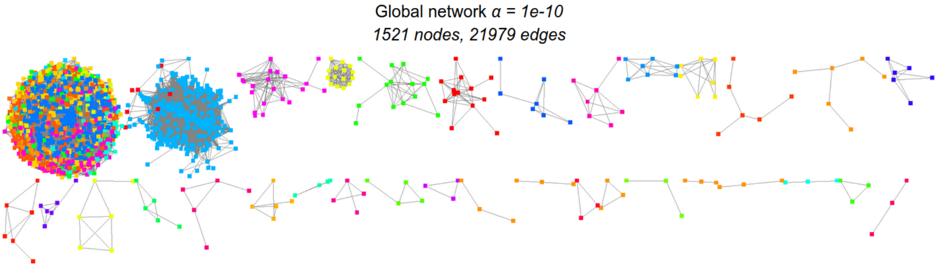


Figure A.2: The network consists of a group of disconnected sub-networks. The largest component (top left) contains two thirds of the total nodes.

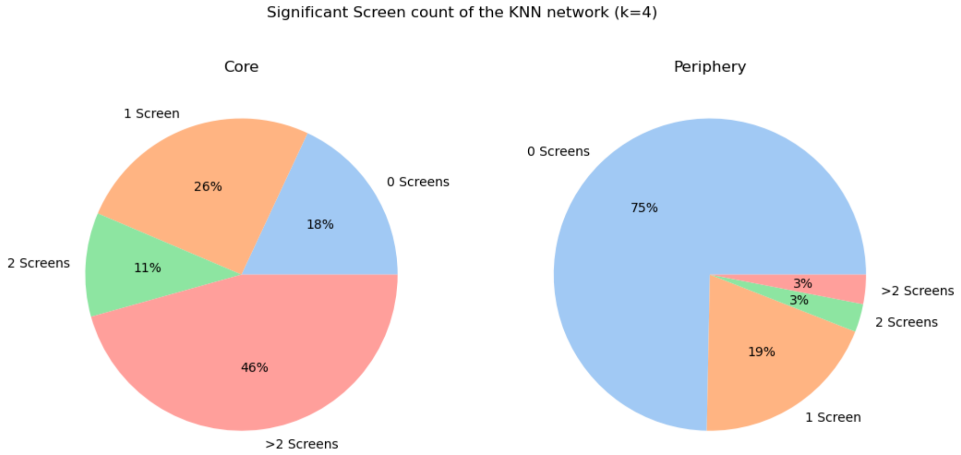


Figure A.3: Pie-chart showing the distribution of genes scoring in 0, 1, 2 or >2 screens.

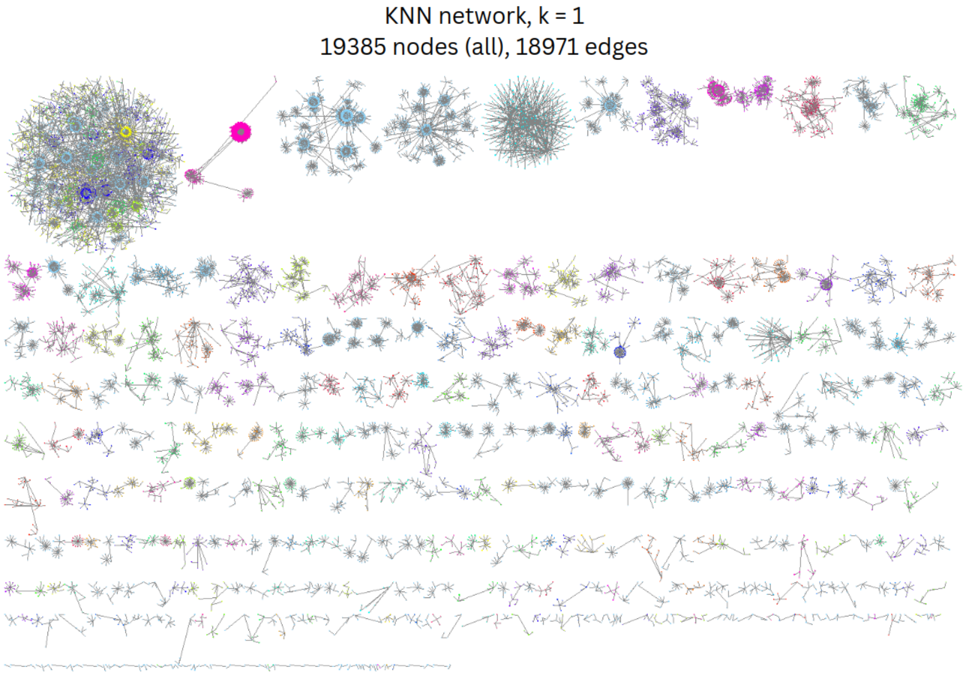


Figure A.4: Overview of the KNN $k = 1$ network.

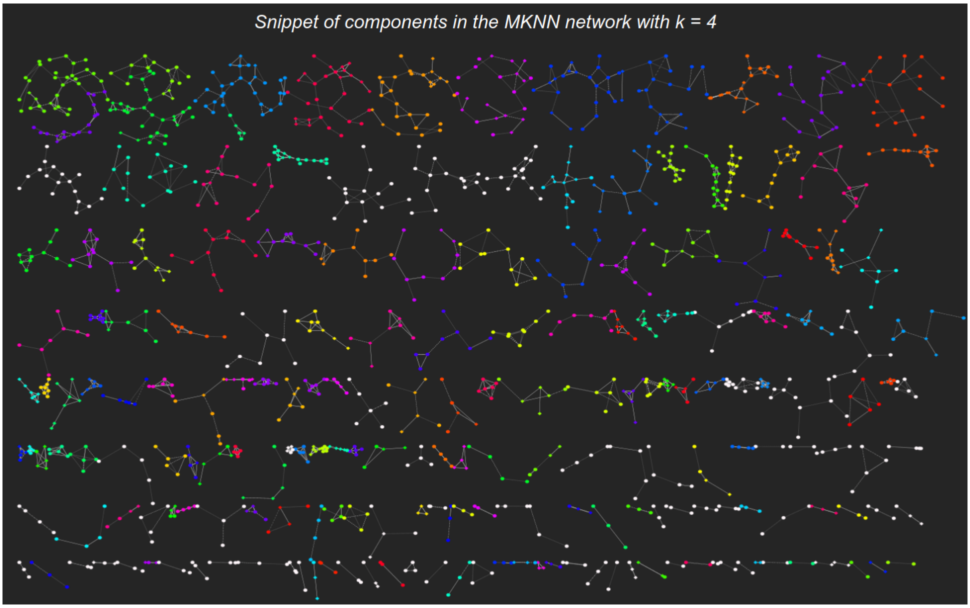


Figure A.5: Snippet of separate components in the MKNN $k = 4$ graph. Nodes are coloured according to unique GO:BP terms, with white nodes representing no enrichment.

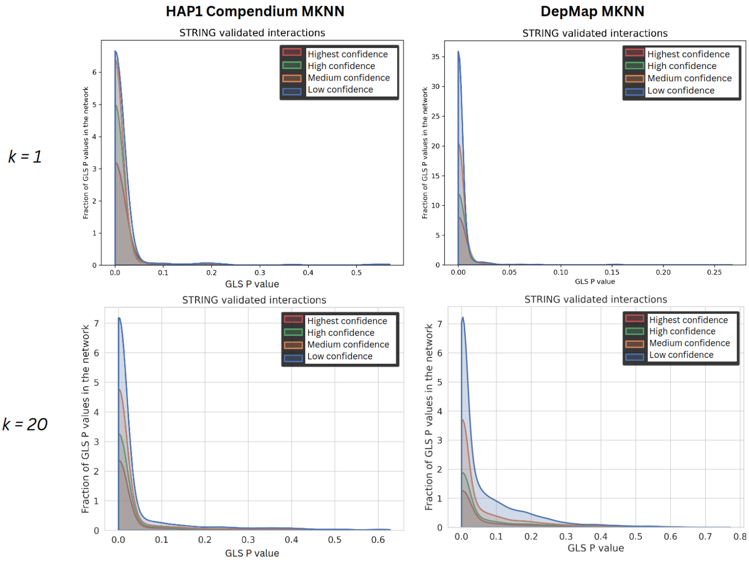


Figure A.6: Density of GLS P value interactions in the STRING database for HAP1 and DepMap at a wide range of k.

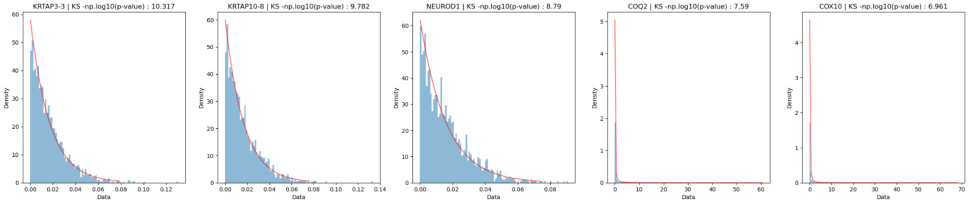


Figure A.7: The five genes with the worst fitting GPD in D_H.

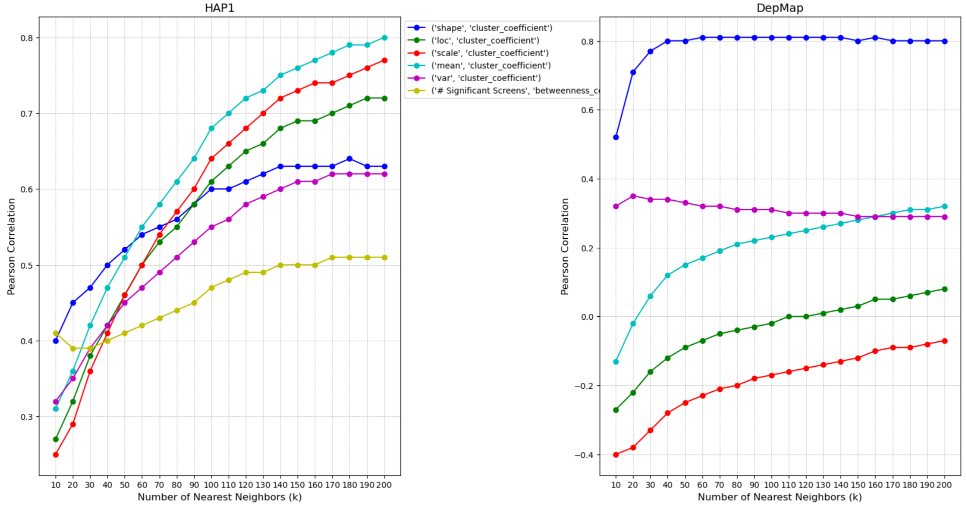


Figure A.8: Pearson correlations of GLS distribution parameters and the clustering coefficient of genes in the MKNN networks for k in range [10, 200].

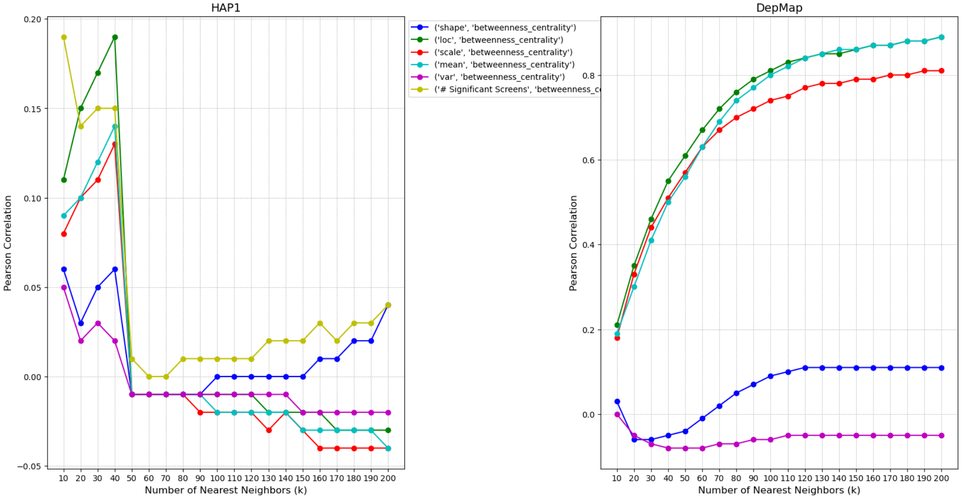


Figure A.9: Pearson correlations of GLS distribution parameters and the betweenness centrality of genes in the MKNN networks for k in range [10, 200].

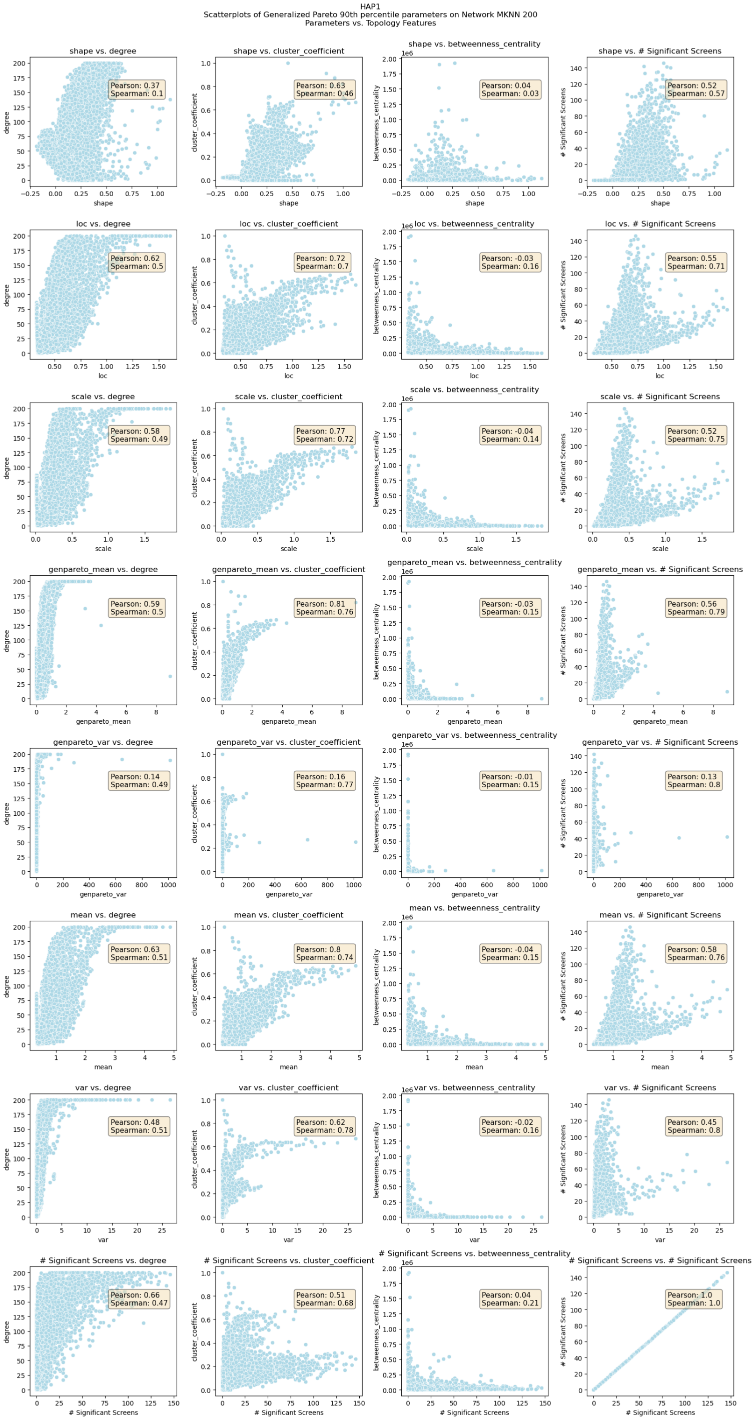


Figure A.10: HAP1: Correlations between three network connectivity parameters and several parameters of the generalized Pareto distributions.

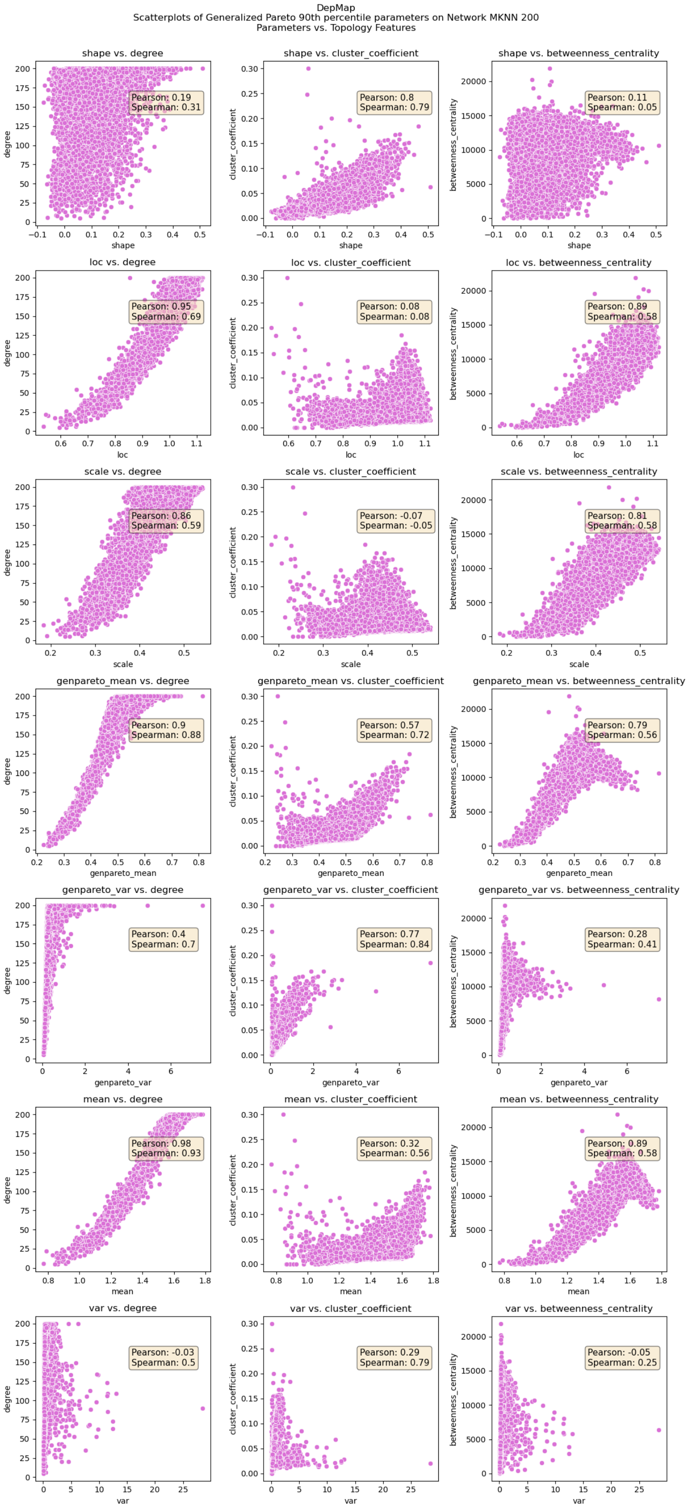


Figure A.11: DepMap: Correlations between three network connectivity parameters and several parameters of the generalized Pareto distributions.

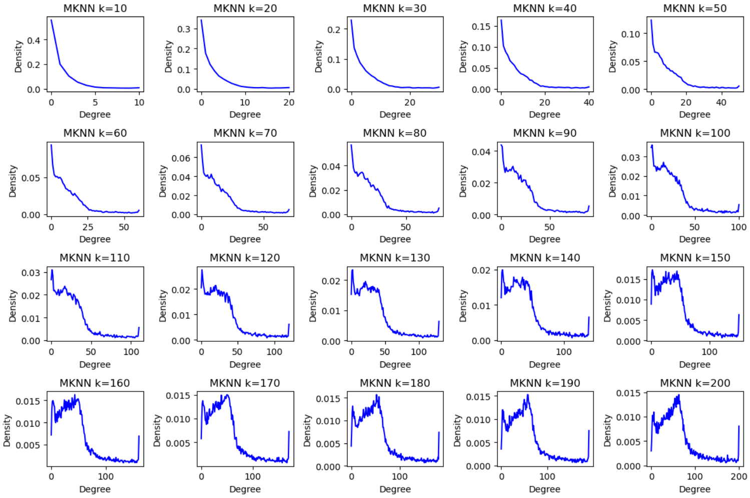


Figure A.12: Degree distributions of the MKNN networks in the HAP1 dataset over a range of *k* values.

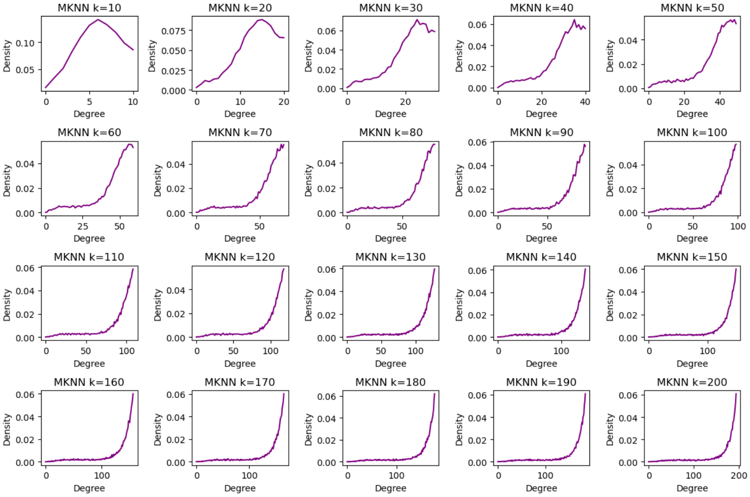


Figure A.13: Degree distributions of the MKNN networks in the DepMap dataset over a range of *k* values.

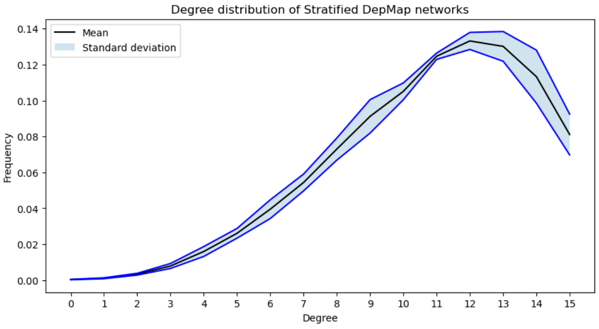


Figure A.14: Degree distribution of the DepMap splits colored for mean and standard deviation over splits.

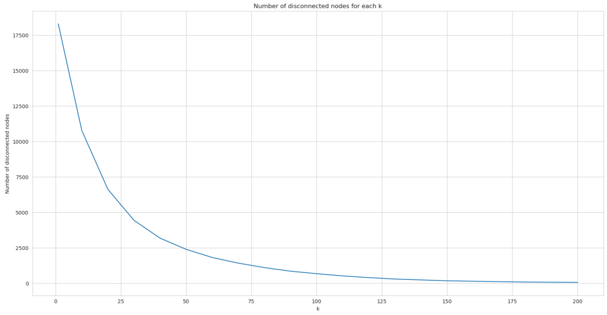


Figure A.15: Number of nodes that have a degree of 0 over a range of KMNN networks.

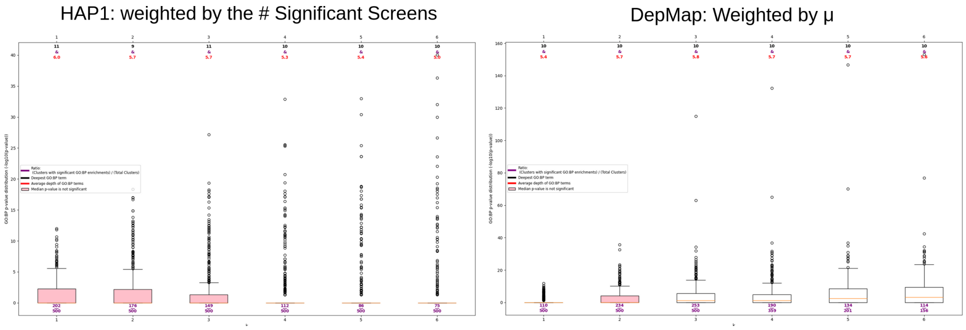


Figure A.16: Distribution of GO:BP enrichment P values of Weighted-MKNN networks for both datasets.

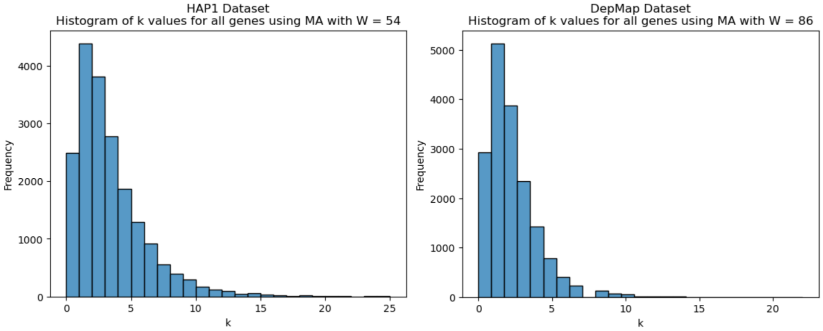


Figure A.17: Distribution of k values for all genes in the datasets based on the number of genes left in the core group using a MA window or 54 or 86 respectively. When no robust area was found using this window, genes are assigned $k = 0$.

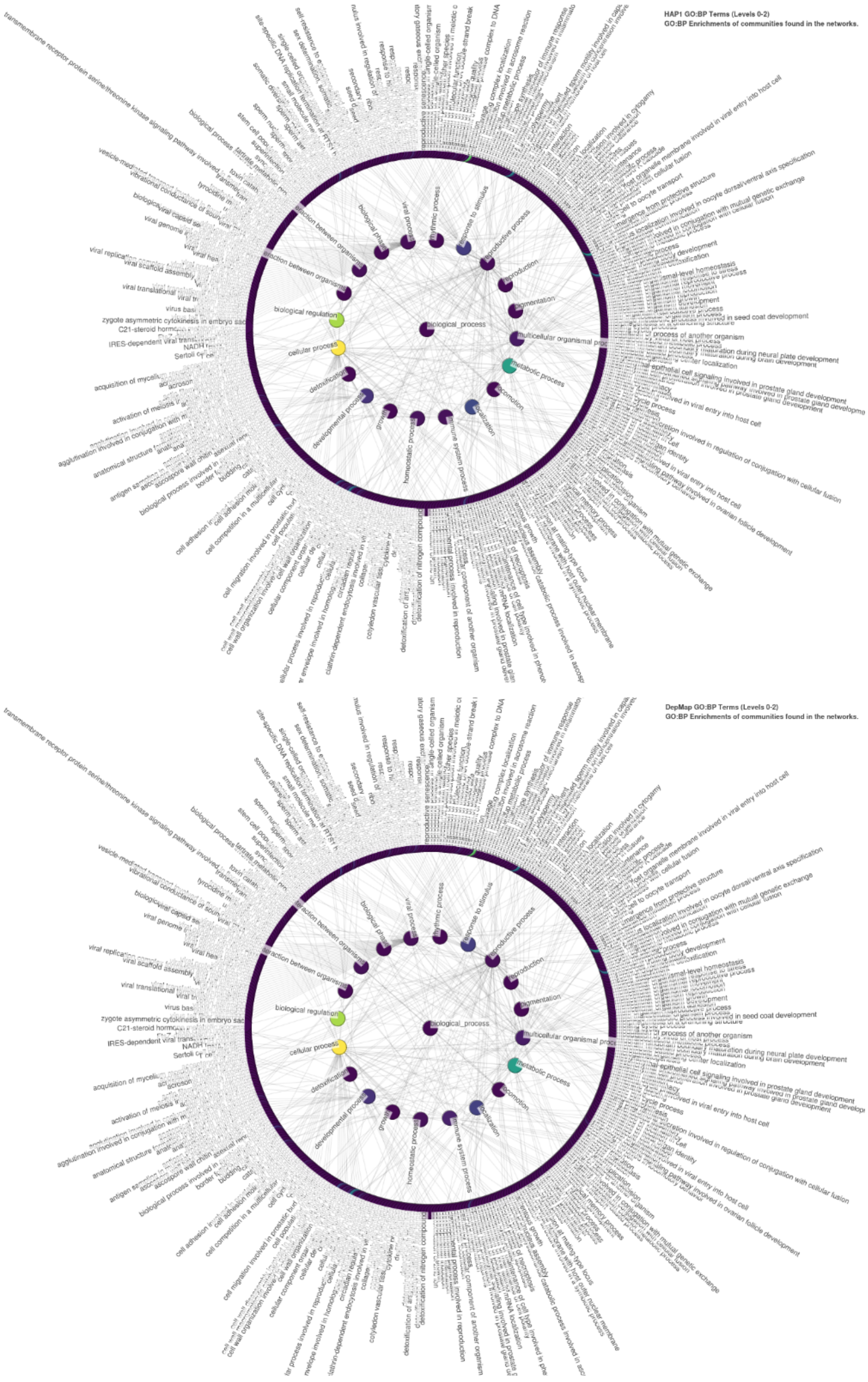


Figure A.18: Coverage of 1st and 2nd level GO:BP terms of HAP1 (top) and DepMap (bottom) datasets.

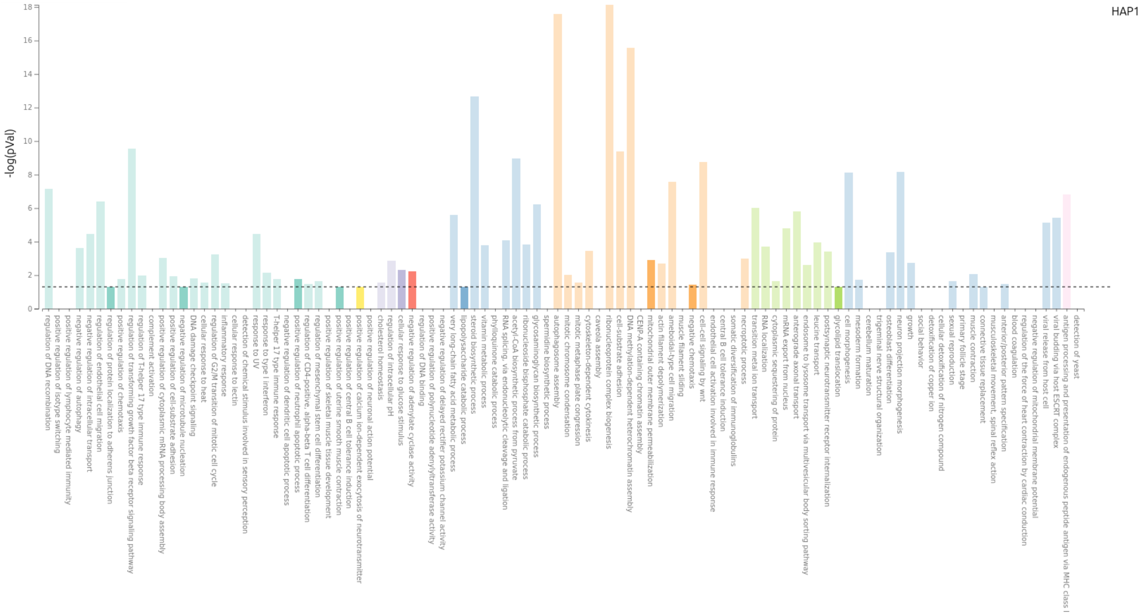


Figure A.19: Bar graph of enrichment -log10(P values) of 100 GO:BP terms found using GO-Compass.

B

Supplementary Tables

Term	Definition
Degree	The number of edges of a vertex to other vertices. The connected vertices are that nodes' neighborhood.
Clustering Coefficient	Measures if a node's neighbors show local clustering. $$C(v) = \frac{\Sigma e_{nv}}{\max(e_{nv})} \tag{B.1}$$ It is the number of edges between neighbors of v divided by the maximum possible amount of edges between neighbors of v.
Betweenness centrality	Measures a nodes' importance of connectivity in the graph. $$B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}. \tag{B.2}$$ σ represent a shortest path between two nodes. Thus, $B(v)$ is the fraction of shortest paths between all node pairs that go through v over all the shortest paths between node pairs.
Neighborhood Connectivity	The average degree of a nodes' neighbors.

Table B.1: Graph Theory Terms and Definitions

Control Cluster Genes	**GO Biological Process Enriched Term**
EED, EZH2, SUZ12, MTF2	Histone H3-K27 methylation
MAPKAP1, RICTOR, INPP4A, MLST8, INPPL1	TOR signalling
VPS16, VPS33A, VPS18, VPS41, VPS11	Endosomal vescicle fusion
WASHC4, COMMD8, VPS35L, CCDC22, COMMD3	Endosomal transport
CD274, PTPN2, IFNGR2, JAK2, IFNGR1	Type II interferon-mediated signaling pathway
DAG1, LARGE1, B4GAT1, RXYLT1, B3GALNT2	Protein O-linked glycolisation
COA3, PDSS2, NARS2, COA7, COX17	Mitochondrial cytochrome c oxidase assembly
ACSL3, HILPDA, GPAT3, ACSL1, DGAT1	Long-chain farry-acyl-CoA metabolic process
SMAD5, BMPR2, SMAD7, SMURF1, SMURF2	BMP signalling pathway
KRAS, NF1, MAPK1, SHOC2	Ras protein signal transduction
CTNNB1, CTNND1, AXIN1	Canonical Wnt signaling pathway
CDK6, SKP2, CDKN2C, CCND3, CDK4	G1/S transition of mitotic cell cycle

Table B.2: **HAP1 Screen Compendium** Control Clusters with associated GO Biological Process Enriched Terms

Control Cluster Genes	**GO Biological Process Enriched Term**
SUZ12, EED, RING1, EZH2, BCOR	Histone modification
MAPKAP1, MLST8, PDPK1, AKT2, AKT1	Positive regulation of protein phosphorylation
VPS18, VPS33A, VPS41, VPS16, FLCN, ARL8B, BORCS5, SPG21, TBC1D5	Endosomal vescicle fusion
WASHC4, COMMD8, VPS35L, CCDC22, COMMD3	Endosomal transport
CDK6, CCND3, TGIF2, MYC, MAX	Negative regulation of transcription by RNA polymerase II
DDX3X, RPS4X	Positive regulation of canonical Wnt signaling pathway
INTS12, INTS6, INTS7, PRCC	snRNA 3'-end processing
KRAS, PRKCE, RAF1, SHOC2, BRAF, MAPK1, CTNNB1, TCF7L2	protein serine kinase activity
MRPL1, RMND1, MPV17L2	Mitochondrial translation
TRIT1, COQ7, COQ6, ATPAF2	Ubiquinone biosynthetic process

Table B.3: **DepMap** Control Clusters with associated GO Biological Process Enriched Terms

C

Details on Implementation for Reproducibility

C.1. ARACNE-AP

The ARACNE-AP package was installed following the instructions on https://github.com/califano-lab/ARACNe-AP. It requires the JAVA JDK > 1.8 and the ANT package. As ANT is not installed on NKI's RHPC cluster, the repository was built with ANT on a local Linux distribution, and the created .jar file then moved to the cluster storage. The program requires two inputs, a gene expression matrix and a list of regulators (e.g. transcription factors). The list of transcription factors (TFs) used was defined by the 2018 Cell paper by Lambert et al., which provides a catalog of likely TFs in human cells [59], and was downloaded directly from http://humantfs.ccbr.utoronto.ca/download.php. The substitute for the gene expression matrix was the previously defined X_H. For the experiment using D_H as input, the matrix was defined as 19385 genes and correspondingly 19385 samples, with the median GLS P values rounded to 6 decimals.

C.2. GO-Compass

GO-Compass uses a React based web-tool for the visualization of results and a Python back end for the implementation of the algorithm. The webtool is available on https://go-compass-tuevis.cs.uni-tuebingen.de/. However, when preparing the data according to their specified directions, an error occured when running the tool. Upon downloading the Github repository, we found this was due to a bug in the React front end which referred to an outdated library. We therefore corrected this mistake and ran GO-Compass locally. This built can of course be shared or forked to a GitHub repository upon request.