January 2020

Towards Personalized Medicine: Computational Approaches For Drug Repurposing And Cell Type Identification

Azam Peyvandipour
Wayne State University

TOWARDS PERSONALIZED MEDICINE:
COMPUTATIONAL APPROACHES FOR DRUG REPURPOSING AND CELL TYPE IDENTIFICATION

by

AZAM PEYVANDIPOUR

DISSERTATION

Submitted to the Graduate School,

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2020

MAJOR: COMPUTER SCIENCE

Approved By:

____

Advisor Date

____

____

____

____

DEDICATION

To my beloved parents for their unlimited love and support.

ACKNOWLEDGEMENTS

First, I would like to express my heartfelt gratitude to my advisor, Prof. Sorin Draghici, for his thoughtful advice, insightful criticism, and wise guidance through the duration of my studies and beyond. He provided me with motivation and inspiration and I thank him for his insights and unending patience during this Ph.D. journey.

I extend my warm thanks to my committee members, Dr. Finley, Dr. Schwiebert, and Dr. Kotov, for providing suggestions, comments and questions that greatly helped in improving the quality of the work. I strongly appreciate their time to review this thesis.

I am also very appreciative of the kind support and constructive advices that I received from Dr. Schwiebert in different steps of my doctoral education.

Next, I would like to thank my family, Maryam, Hojjat, Azadeh, Mahdi, and Amir for their unwavering love and support. My beloved mother, Maryam, was and remains my best role model for hard working and my first teacher. My father, Hojjat, for teaching me the significant values of life. Thanks to my sister and brother, Azadeh and Amir, for showing immense love and care and always motivating me in my career. Finally, I thank my brother-in law, Mahdi for being unconditionally understanding and encouraging.

I also would like to thank all the past and current members in ISBL team for their encouragement and support. Specially, I like to thank Dr. Michele Donato, Dr. Cristina Mitrea, Dr. Sahar Ansari, Nafiseh Saberian, and Dr. Adib Shafi for their great collaboration and friendship. I also thank my friends in Wayne State University, Hedieh Torabifard, Mitra Asgari, and Sanaz Rabinia who were supportive and helping me toward my Ph.D. Finally, I extend my thanks to all people who were involved in the process. I am also grateful for all the financial support from the Computer Science Department, Wayne State University.

This research was supported in part by the following grants: NIH/NIDDK R01 107666A, National Science Foundation (SBIR 1853207), and W81XWH-16-1-0516.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

Despite enormous investments in research and development (R&D), it still takes approximately $800 million to $2 billion and 10-17 years to approve a new drug for clinical use [5, 67, 68]. More than 90% of drugs fail to pass beyond the early stage of development and toxicity tests, and many of the drugs that go through early phases of the clinical trials fail because of adverse reactions, side effects, or lack of efficiency. Based on a recent report [244], around 90% of drugs are effective only on 40% of patients, and such ineffective treatment can cause an enormous loss ($350 billion/year) in the United States alone.

We are addressing above mentioned challenges by introducing two main strategies. First, we are developing a novel computational method to discover novel therapeutic roles for existing FDA-approved drugs. The identification of novel disease indications for approved drugs, known as drug repositioning (or repurposing), is a very effective way to increase the therapeutic arsenal at a very reduced cost [16, 42, 54, 173, 205, 232, 238, 239, 292]. Finding new disease indications for existing drugs sidesteps all these issues and can therefore increase the available therapeutic choices at a fraction of the cost of a new drug development. The need for new drugs is currently met mostly with the classical drug development pipeline, which is slow, extremely expensive, and very prone to failures. Drug repurposing tools are likely to be widely adopted by pharmaceutical companies because they would offer a very low cost solution to increase the market size for their existing drugs [167, 237]. The societal impact includes increasing the available therapeutic choices for existing diseases, reducing treatment cost by offering more alternatives, as well as possibly finding treatments for “orphan diseases” - rare diseases that will never have targeted drugs because they do not justify the usual drug development cost for their market size.

In this thesis, new usages for existing FDA-approved drugs are identified by performing a system-level analysis using gene expression data and publicly available data, including drug targets, disease-associated genes and KEGG signaling pathways.

Analyzing the diversity and evolution of single cancer cells can also enable the advances in early cancer diagnosis, and ultimately choosing the best strategy for cancer treatment [149, 231, 244]. In particular, this analysis can play a crucial role in cancer treatment, where individual cells develop drug resistance and metastasis [122, 136, 244].

Recent advances in single-cell RNA-Seq (scRNASeq) techniques have provided transcriptomes of the large numbers of individual cells (single-cell gene expression data) [61, 79, 99, 126, 168, 186, 194, 266, 280]. Unlike the bulk measurements that average the gene expressions over the individual cells, gene measurements at individual cells can be used to study several different tissues and organs at different stages [99, 126, 186, 266, 280]. Furthermore, one important analysis on scRNASeq is the identification of cell types that can be achieved by performing an unsupervised clustering method on transcriptome data [12, 13, 75, 177, 291, 304, 308].

In this thesis, as a second strategy, we have introduced a novel method to identify the cell types using single-cell gene expression data. To do this, we have developed a pipeline to cluster the individual cells based on their gene expression values such that each cluster consisting of cells with specific functions or distinct developmental stages. We used the Adjusted Rand Index (ARI) [113], adjusted mutual information (AMI) [274, 275], and V-measure [226] to evaluate the performance of the clustering result for datasets in which the true cell types are known.

This thesis is organized as follows. Chapter 2 focuses on the proposed approach for drug repurposing, introduced in the context of the systems biology. In this approach, new usages for existing FDA-approved drugs are identified by performing a system-level analysis using gene expression data and publicly available data, including drug targets, disease-associated genes and KEGG signaling pathways. The proposed approach first builds a drug-disease network (DDN) by considering all interactions between drug targets and disease-related genes in the context of signaling pathways [205]. This network is integrated with gene-expression measurements to identify drugs with new desired therapeutic effects. This method is evaluated based on its ability to re-discover drugs that are already FDA-approved for a given disease.

Our approach is innovative because it focuses on the effect at the pathway level, rather than the effect on a specific set of genes. Cancer could be used as a good high-level example [63]. One of the “hallmarks” of solid tumors independently of the type of cancer or localization is that the cancer cells manage to avoid apoptotic signals. In other words, their apoptotic pathway is inhibited. The specific set of genes that each particular type of tumor uses to accomplish this is less relevant than the ultimate effect, which is uncontrolled proliferation. Identifying the specific genes that are dis-regulated in a given type of tumor and then looking for drugs that have an antagonistic effect on the very same set of genes, as proposed by Sirota et al. [248] is a sufficient but perhaps not necessary condition. Identifying a drug that has an antagonistic effect on the apoptotic pathway may be sufficient, even though this drug may induce its effect through a different set of genes than that used by the tumor to escape apoptosis in the first place.

Chapter 3 focuses on recent methods for the identification of cell types based on unsupervised clustering of the single cell gene expression data. In this chapter, we propose a framework that addresses challenges in identifying meaningful clusters of cells. We validate the performance of the proposed method on eight publicly available scRNA-seq datasets with known cell types as well as five simulation datasets with different degrees of the cluster separability. We compare the proposed method with five other existing methods: RaceID [93], SC3 [139], SINCERA [95], SEURAT [162], and SNN-Cliq [291]. The results show that the proposed method performs better than the existing methods. Finally, chapter 4 concludes the dissertation by proposing future work and possible research directions.

CHAPTER 2 DRUG REPURPOSING

2.1 Problem statement

Despite enormous investments in research and development (R&D), it still takes approximately $800 million to $2 billion and 10-17 years to approve a new drug for clinical use [5, 67, 68]. More than 90% of drugs fail to pass beyond the early stage of development and toxicity tests, and many of the drugs that go through early phases of the clinical trials fail because of adverse reactions, side effects, or lack of efficiency. Indeed, the rate of failure is still significantly higher than the rate of approval [31, 67, 68]. In order to overcome these challenges, drug repurposing, an approach aiming to find new indications for existing drugs [54], has emerged as an important strategy for drug discovery [16]. This approach can also rescue drugs that are safe but fail to get to market due to the lack of efficacy against their initial clinical indication [58]. A well-known example of repurposed drug is Thalidomide. Originally, this drug was approved for the treatment of morning sickness during pregnancy. Not long after that, it was withdrawn from the market due to the severe birth defects in the early 1960s [135]. However, several years after, it was approved by the FDA for the treatment of multiple myeloma [20].

2.2 Overview of existing approaches

Repurposing approaches can be categorized as drug-based or disease-based. Disease-based approaches are developed to overcome the lack of knowledge about the pharmacology of a drug [41, 74]. Drug-based approaches are preferred when drug data (e.g. transcriptomic data) are available. While each one of these approaches faces several challenges, successful repurposing approaches often take advantage of both drug and disease data. In this area, a number of of approaches have been developed based on the analysis of transcriptomic data, such as gene expression signatures, defined as the changes in the expression of genes under a certain condition (e.g. administration of a drug, or a disease). Some of these approaches are based on the idea that if there is an anti-correlation between a drug-exposure gene expression signature and a disease gene expression signature, that drug may have a potential therapeutic effect on the disease [146, 248]. Drugs that are strongly anti-correlated with a disease are likely to be candidates for repurposing. Resources such as LINCS (new version of Connectivity Map [146]) allow for systematic search of candidates for drug repurposing.

The Connectivity Map (CMap) project [146] was the first systematic approach aimed at exploring functional connections between drugs, as well as between drugs and diseases. This project led to the first repository of genome-wide expression data from five human cancer cell lines exposed with 1,309 compounds at different dosages, and integrated with other sources such as NCBI Gene Expression Omnibus (GEO). [146] evaluate the similarity of a query signature, that can be a drug-exposure gene expression signature or a disease gene expression signature, to each drug signature in Connectivity Map database (reference data). In [248] the authors developed a systematic approach based on the same idea originally proposed by [146]. In this work, they use drug-exposure gene expression signature from Connectivity Map as the reference data and query this reference data with every single disease gene expression signature by applying a pattern-matching method.

Some approaches [49, 116, 198] employ Over-Representation Analysis (ORA) in order to understand the mode of actions (MoA) of drugs and their potential new usages. The most recent approach [116], DGE-NET, is based on the hypothesis that drugs with similar binding patterns (to a reference target protein) have similar molecular activities. In step

1, drug-target interactions (drug-target signatures) are predicted. DGE-NET ranks drugs that most likely bind to a given target based on similarity scores between different drugs and the given target. To do this, a modified version of Train, Match, Fit and Streamline (TMFS) [62] is used. TMFS determines the binding potential of a protein-ligand complex incorporating docking, three-dimensional shape, and ligand physicochemical data. For a given protein target, the top 40 drugs (1% of all drugs) are selected as hits for the next step. In step 2, the associations between the drug-target signatures on the one hand, and diseases, pathways, functions, and protein-protein interactions on the other hand, are identified. To this end, a hypergeometric test is applied at different biological levels: protein targets, protein-protein interactions (PPIs), cell signaling pathways, and molecular functions. In this step, gene expression data of the given disease is exploited to identify differentially expressed genes by comparing the expression values of two groups of samples: normal and disease. DAVID [110, 111] and STRING analysis [80] is applied on the list of differentially expressed genes to indicate the associations at different levels (by computing z-scores). The identified associations are validated based on the current literatures and annotated databases. DGE-NET is applied to human disease gene expression datasets: rheumatoid arthritis, inflammatory bowel disease, Alzheimer's disease, and Parkinson's disease to prioritize FDA-approved drugs for repurposing purposes.

Another approach [198] performs the pathway enrichment analysis to investigate MoA and clinical functions of the FDA-approved drugs. First, it selects sixteen FDA-approved drugs (with available target information) from DrugBank [287]. Then, it retrieves primary and secondary targets of these drugs from GEO datasets [76], MMDB [164], and PubChem [153]. And finally, it applies the enrichment analysis based on a modified Fisher's

Test using the drug targets and pathways data. Pathways are ranked based on the numbers of retrieved drug targets involved in each pathway. Pathways with p-values <0.05 are chosen for further investigations.

The third method using an ORA approach analyzes the associations among drugs, targets and biological functions [49]. It assigns drugs into nine classes based on their targets: (1) G protein-coupled receptors, (2) cytokine receptors, (3) nuclear receptors, (4) ion channels, (5) transporters, (6) enzymes, (7) protein kinases, (8) cellular antigens and (9) pathogens. Then, it employs an enrichment analysis to identify the associations between the drugs and features including GO terms [17] and KEGG (`http://www.genome.ad.jp/kegg/`) pathways. Thus, given the drug and the KEGG pathway (or GO term), an enrichment score is computed as S_KEGG (or S_GO) based on the result of hypergeometric test (p-value). Overall, 279 KEGG pathways and 17,904 GO terms are exploited to obtain the enrichments scores. Each drug is represented by 279 S_KEGG enrichment scores and S_GO 17,904 enrichment scores. Finally, the feature selection minimum redundancy maximum relevance (mRMR) method [203] is used to extract the key features. Pathways and GO terms that are highly enriched by several classes of drugs can be investigated for drug interaction predictions. For instance, the neuroactive ligand-receptor interaction pathway is enriched with two classes: GPCR and IC. This suggests that drugs with different targets may belong to the same biological pathway, thus also suggesting a potential for synergistic drug interactions.

The lack of a unifying analysis at system-level makes such ORA methods limited. In this study, we use: i) KEGG signaling pathways, ii) drug target data, and iii) disease associated genes to construct a global network with genes specific to the drug and disease of interest (called drug-disease network, DDN). We then measure gene perturbation signatures for drug-disease pairs by propagating measured expression changes across the network topology.

Methods based on protein-protein interactions that employ over-representation analysis (ORA) are limited by the fact that each gene is analyzed independently without a unifying analysis at a system level, while the proposed approach aims to consider the system-level dependencies and interactions on the drug-disease-specific network. One of the existing approaches for drug repurposing is based on an over-representation analysis of various pathways, based on the genes targeted by a given drug [116]. Based on this approach, a drug is first associated with pathways based on its directly targeted genes. Using this approach, Sunitinib can be associated to the following KEGG pathways: *MAPK signaling pathway*, *Cytokine-cytokine receptor interaction*, *VEGF signaling pathway*, and *Pathways in cancer*. Subsequently, Issa *et al.* expand the set of genes to include the genes having direct interactions with the Sunitinib target genes using PPI data, and recalculated the pathways enriched in these “predicted targets”. Table 1 shows the list of pathways that are significantly enriched in such predicted targets (FDR-corrected p-values less than 0.05).

Although this type of ORA analysis could provide useful associations between Sunitinib and various pathways, extrapolating from pathways to diseases may not be optimal since there are many diseases that might be relevant to these pathways. For instance, the KEGG pathway *Pathways in cancer* is associated with over 1,000 diseases according to CTD [169]. This illustrates why this simple pathway-based enrichment approach cannot be used for effective drug repurposing and explains why our system-level analysis is able to provide much more specific results.

Table 1: A list of pathways that are significantly enriched in predicted target genes for Sunitinib (FDR-corrected p-value < 0.05).

Pathway	pORA.fdr
PI3K-Akt signaling pathway	1.08E-18
Cytokine-cytokine receptor interaction	1.11E-15
Focal adhesion	2.80E-14
Pathways in cancer	1.20E-12
Melanoma	1.75E-07
Prostate cancer	8.30E-07
mTOR signaling pathway	1.08E-06
Gap junction	1.29E-05
MAPK signaling pathway	3.73E-05
Glioma	3.73E-05
HIF-1 signaling pathway	3.73E-05
Endocytosis	0.00088
Pertussis	0.00093
Regulation of actin cytoskeleton	0.00093
Rheumatoid arthritis	0.00174
Transcriptional misregulation in cancer	0.00317
Amoebiasis	0.00468
Legionellosis	0.00638
Acute myeloid leukemia	0.00747
p53 signaling pathway	0.00904
Long-term potentiation	0.00943
Hepatitis B	0.00982
HTLV-I infection	0.01044
Hypertrophic cardiomyopathy (HCM)	0.01341
Progesterone-mediated oocyte maturation	0.01341
Calcium signaling pathway	0.01605
Oocyte meiosis	0.02104
Osteoclast differentiation	0.04606
Amyotrophic lateral sclerosis (ALS)	0.04606
Insulin signaling pathway	0.04798

In summary, the limitations of these methods can be summarized as follows: i) the obtained target proteins might not be the exact target of a drug (for instance, they can be indirectly associated to the drug because of down or upstream proteins or cross-talk effects); ii) the target(s) of a drug have a limited ability to identify the pathways based on enrichment alone since one or a few genes on a pathway are unlikely to constitute sufficient statistical evidence; iii) the drug targets might be involved in variety of pathways that are not specifically related to the drug primary target. Thus such methods may fail in identifying significant pathways related to biological functions of the drug based on their target information.

Although existing drug repurposing methods showed moderate success, they are far from bringing critical advancements in the drug development pipeline. Most of these approaches rely only on an analysis of a set of differentially expressed genes. However, changes in genes expression are propagated in the system through a complex gene signaling network and this fact is not captured by approaches using only lists of DE genes [72, 133, 180].

It has been shown that many drugs exert their effect through modulation of several proteins rather than single targets [105, 106, 196, 217]. Furthermore, the analysis by [300] shows that not all drugs directly impact the proteins associated with the root cause of a disease. These findings suggest that drug repurposing may be more successful if it used novel paradigms, going beyond lists of genes.

Systems biology can be used as an effective platform in drug discovery and development by leveraging the understanding of interactions between the different system components [3, 14, 34, 60, 141, 205, 259, 260, 270, 297]. In this work, we propose a systems biology approach that takes advantage of prior knowledge of drug targets, disease-related genes, and signaling pathways to construct a drug-disease network (DDN) composed of the genes that are most likely perturbed by a drug [205]. By performing a system-level analysis on this network using disease gene expression signatures and drug-exposure gene expression signatures, our approach estimates the amount of perturbation caused by a drug on the genes that are associated to a disease of interest. Drugs are ranked based on the amount of perturbation they exercise on specific disease-related genes, and highest ranking drugs are proposed as candidates for repurposing.

We compare the results of our approach with the computational drug-repurposing approach proposed by [248] using 19 datasets involving 4 diseases: idiopathic pulmonary fibrosis (IPF), non-small cell lung cancer (NSCLC). We show that our approach provides a more accurate prediction based on its ability to identify drugs that are already approved for the disease of interest.

2.3 Drug repurposing using systems biology

2.3.1 Disease and drug gene expression data

Large scale drug-exposure gene expression data are obtained from two databases: Connectivity Map and the Library of Integrated Network-Based Cellular Signatures (LINCS) [146] (`http://www.lincsproject.org/`).

Disease expression data are obtained from NCBI Gene Expression Omnibus (GEO) [76] and Lung Genomics Research Consortium (`http://www.lung-genomics.org`).

In Connectivity Map, drug expression data are measured from the exposure of 5 human cell lines to bioactive small molecules. Differentially expressed genes (DEGs) are identified using a moderated t-test [250] by comparing treated samples and the corresponding control (untreated) samples. The resulting p-values are FDR adjusted [24] to correct for multiple comparisons.

The LINCS program, the successor of CMap [146], generated transcriptional gene expression data from cultured human cells exposed to small molecules and knock-down-overexpression of a single gene. The data is also available in GEO (GSE70138). This program provides DEGs in terms of z-score signatures by comparing two groups of samples (treatment vs control). Since measurements are carried out on different platforms, we standardize gene identifiers from chip specific probe identifiers to NCBI GeneID identifiers using the affy package [83]. We average across distinct probe expression values when multiple probes mapped to the same NCBI GeneID.

In both Connectivity Map and LINCS, there are often more than one replicate for each drug. Replicates with at least (1%) DEGs (FDR-adjusted p-value <0.025) are selected. Since measurements are carried out on different platforms, we standardize gene identifiers from chip specific probe identifiers to NCBI GeneID identifiers using the affy package [83]. We average across distinct probe expression values when multiple probes mapped to the same NCBI GeneID.

We use two sources: i) Connectivity map (CMap) [146] for breast cancer and prostate cancer, and ii) NIH's Library of Integrated Network-based Cellular Signatures (LINCS) for idiopathic pulmonary fibrosis (IPF) and non-small cell lung cancer. We used two different data sources in order to show the approach is reliable and works independently of the source of the drug data. CMap database has several of FDA-approved drugs for breast cancer and prostate cancer but none for idiopathic pulmonary disease (IPF). So, for IPF we used LINCS database that has several gene expression profiles for Nintedanib (an FDA-approved drug for IPF). We chose to use LINCS for non-small cancer (NSCLC) rather than CMap for two reasons. First, more FDA-approved drugs belong to this database. (four drug instances for each of FDA-approved drugs: Gefitinib and Crizotinib). In comparison, the only FDA-approved drug for NSCLC in CMap is Paclitaxel and there is only one instance for this drug. Second, in LINCS, 123 out of 260 drug instances are measured on the A549 cell line that is the human lung adenocarcinoma epithelial cell line (a model For NSCLC). Since IPF is a chronic progressive and ultimately fatal disease that did not have any effective treatment until recently, we were interested to see if our proposed approach is able to predict any new treatments for this disease.

2.3.2 Drug-targets and disease-related genes

The proposed approach needs to construct a network that includes all the shortest paths between the drug targets and genes known to be associated to the disease of interest. Drug targets and disease-related genes (genes associated with the disease of interest) are retrieved from the Comparative Toxicogenomics Database (CTD) [169] and Drugbank [287]. CTD is a database that provides curated data describing cross-species chemical-gene/protein interactions and gene-disease associations. Drugs with no known targets are removed from the study. Such drugs are mostly not FDA-approved.

2.3.3 Signaling pathways

We obtain signaling pathways from Kyoto Encyclopedia of Genes Genomics (KEGG) (`http://www.genome.ad.jp/kegg/`). A signaling pathway in KEGG is modeled by a graph in which nodes represent genes or proteins, and directed edges between them represent signals between genes or proteins. The edges are weighted based on the various types of signals, such as activation, inhibition, etc.

2.4 Proposed framework

2.4.1 Drug-disease network construction

The first part of the framework consists in building the drug-disease network (DDN) by integrating knowledge about the disease-related genes, drug targets, and gene-gene interaction knowledge. Then, a repurposing score is computed for each drug-disease pair by integrating expression data into this network. Figure 1 represents the proposed framework. As shown in Figure 1A, first, we construct a global network (GN) by performing the union of all nodes and edges of KEGG human signaling pathways. In a number of KEGG pathways, a gene ’a’ interacts with gene ’b’, through an intermediate pathway ’A’. This is represented by a link that starts from gene ’a’ to gene ’b’ through pathway ’A’. For example, in the Adherence Junction pathway, TGFβR activates Smad3 through the TGF-beta signaling pathway. Interactions between genes belonging to the pathway ’A’ and genes ’a’ and ’b’ are not included in our model. There are some interactions between genes/pathways through DNA or small molecules in KEGG. For instance, there is a link between MAPK signaling pathway and Phosphatidylinositol signaling system through a small molecule (compound) IP3 in KEGG. Such interactions are not part of the scope of this analysis and we do not include them in constructing the global network [205]. We used ROntoTools package [276] (version 1.2.0) to calculate the union all KEGG signaling pathways that are represented by the adjacency matrices and obtain a unified adjacency matrix. In this step, we included some implicit interactions between the genes by performing the union of adjacency matrices representing KEGG signaling pathways. For example, suppose gene ’a’ activate gene ’b’ in pathway ’A’ and gene ’b’ activates gene ’c’ in pathway ’B’. A path between gene ’a’ and gene ’c’ may be constructed by our analysis, while there was no path between them before this analysis.

Next, given the two sets of disease-related genes as $Disease_t = \{x_1, x_2, ..., x_n\}$, and drug targets as $Drug_t = \{y_1, y_2, ..., y_n\}$, we extract a subgraph of GN that consists of all the shortest paths connecting genes belonging to these sets. It means that a gene from either $Disease_t$ or $Drug_t$ can be a source or destination of the shortest path extracted from GN. This subgraph called Drug-disease network (DDN) represents all the interactions between drug targets and genes related to the given disease, through all the interactions

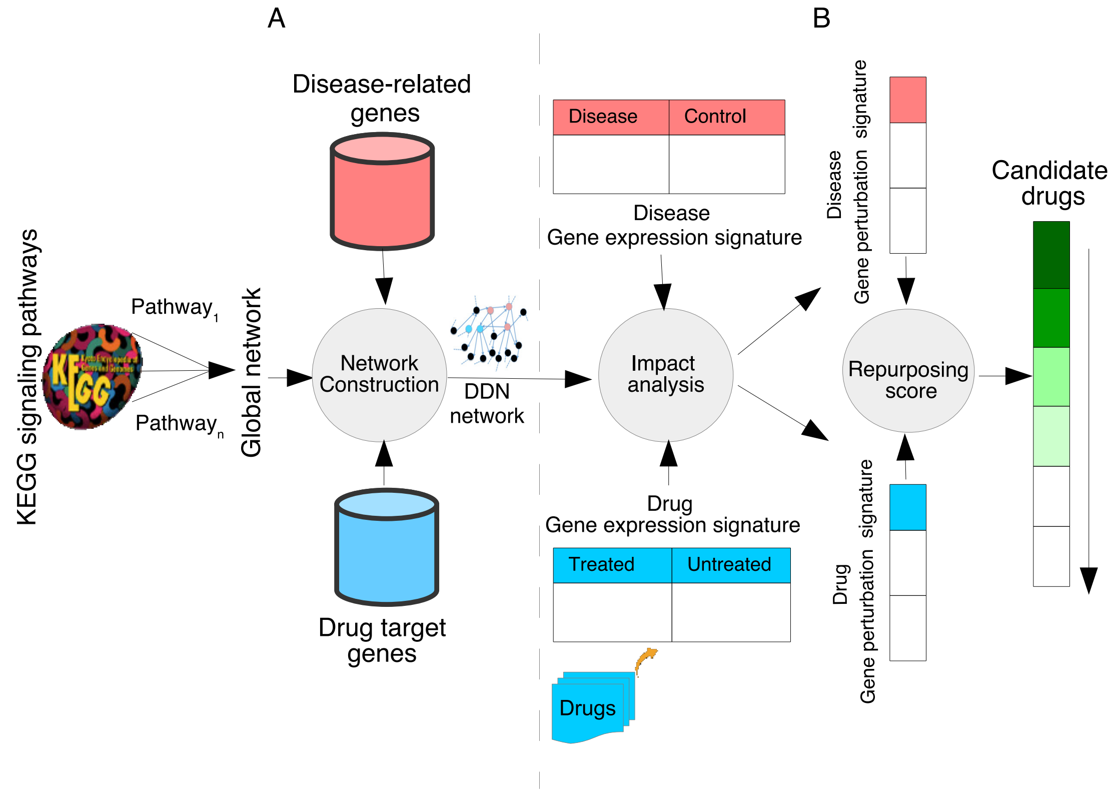


Figure 1: Framework overview. A) We construct a global network (GN) that is the union of all KEGG human signaling pathways. For each drug-disease pair, we extract a subgraph of *GN*, namely DDN, consisting of all shortest paths between two sets of disease-related genes and drug targets. B) We then generate gene perturbation signatures of drug-disease pairs by applying a system-level analysis on their gene expression signatures in the drug-disease network (DDN). A comparative analysis is applied on drug and disease gene perturbation signatures. A repurposing score is assigned to each drug-disease pair. Finally, a ranked list of drugs with potential therapeutic effects for the given disease is generated based on repurposing scores.

described in KEGG signaling pathways.

2.4.2 Drug-disease repurposing score computation

In this stage, we capture the impact caused by a drug exposure or a disease on the genes that are specific to the condition of interest. In order to integrate the drug and disease gene expressions signatures, we generate gene perturbation signatures by computing the amount of perturbation upon the genes belonging to the drug-disease network (DDN) for all drug-disease pairs, as shown in Figure 1B. The gene perturbation signatures are calculated using the impact analysis method [71] on the subgraph of global network we constructed in previous step. The impact analysis (IA) takes into account the structure and dynamics of a signaling pathway by considering a number of important aspects, including the measured gene expression changes, the direction and type of every gene signal, and the position and role of every gene in a pathway. A perturbation factor for each gene, $PF(g_i)$, is calculated using the impact analysis method [71], as follows:

A perturbation factor for each gene, $PF(g_i)$, is calculated using the impact analysis method [71], as follows:

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \frac{PF(g_j)}{N_{ds}(g_j)} \tag{2.1}$$

where the term $\Delta E(g_i)$ denotes the signed normalized measured expression change of a gene g_i, added to the sum of all perturbation factors of the genes g_j that are direct upstream of the gene g_i, normalized by the number of downstream genes of g_j, $N_{ds}(g_j)$. The coefficient β_{ij} represents the type of the interaction, $\beta_{ij} = 1$ for activation and induction, and $\beta_{ij} = -1$ for inhibition and repression. The second term in Equation (2.1) involves the PF values of those genes that are upstream of the gene for which the perturbation factor is calculated. For a gene with no upstream genes, the PF will be the measured expression gene $\Delta E(g)$.

Next, we calculate the repurposing scores for drug-disease pairs by computing the Pearson correlation coefficient between their gene perturbation signatures. The result score is from -1 to 1, where a high positive score shows that the drug and the disease both cause similar perturbations in the system, and therefore, that drug may cause the same effect as the disease. Conversely, a high negative score shows that the drug and disease have opposite gene perturbation signatures. Our hypothesis is that if the perturbation caused by a particular drug in the system is the reverse of the perturbation caused by a disease, that drug may have the potential to treat the given disease. Thus, we rank drugs from the strongly anti-correlated to the strongly correlated, according to their repurposing pathway perturbation scores.

In order to estimate the statistical significance of drug candidate repurposing scores, we generate 1,000 random drug gene expression signatures (by permuting gene labels), and then calculate random repurposing scores for all drug-disease pairs. We compute *p*-values as the percentage of the random scores higher than the observed score.

2.4.3 A systematic method to select repurposing candidates

We used a systematic method in order to rank repurposing candidates. To do this, given a ranked-list of drugs (drug instances) obtained by applying our approach on a disease dataset, we first compute a score for each drug that indicates how better or worse that drug is ranked in comparison to already FDA-approved drugs as follows:

$$score(Drug_x) = a - b \tag{2.2}$$

where a denotes the number of already FDA-approved drugs (gold standards) that are ranked worse than $Drug_x$, and b denotes the number of FDA-approved drugs that are ranked better than $Drug_x$ (see Figure 3). For instance, if there were N FDA-approved drugs for a condition and an instance of a repurposing candidate were ranked higher than all N FDA approved drugs, the score of this candidate would be N. Conversely, if the candidate were ranked lower than all N FDA approved drugs, its score would be $-N$.

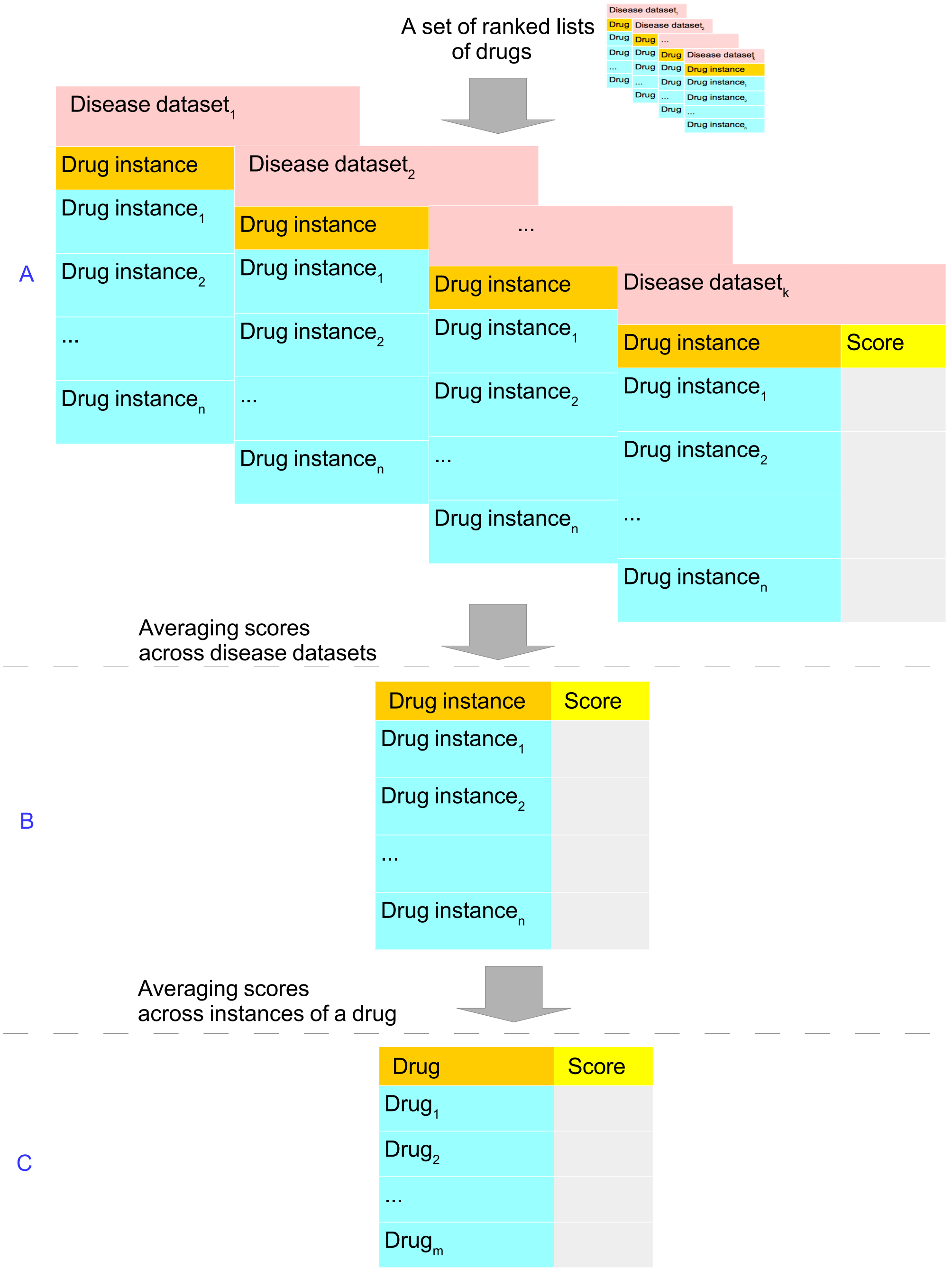


Figure 2: A systematic method [205] to select repurposing drug candidates. A) Given a set of ranked lists of drugs (drug instances), we compute a score for each drug. B) We then calculate an average score for each drug instance across different lists (using disease datasets). C) Finally, we calculate an average score for each distinct drug across the instances, in case there are multiple instances for that drug.

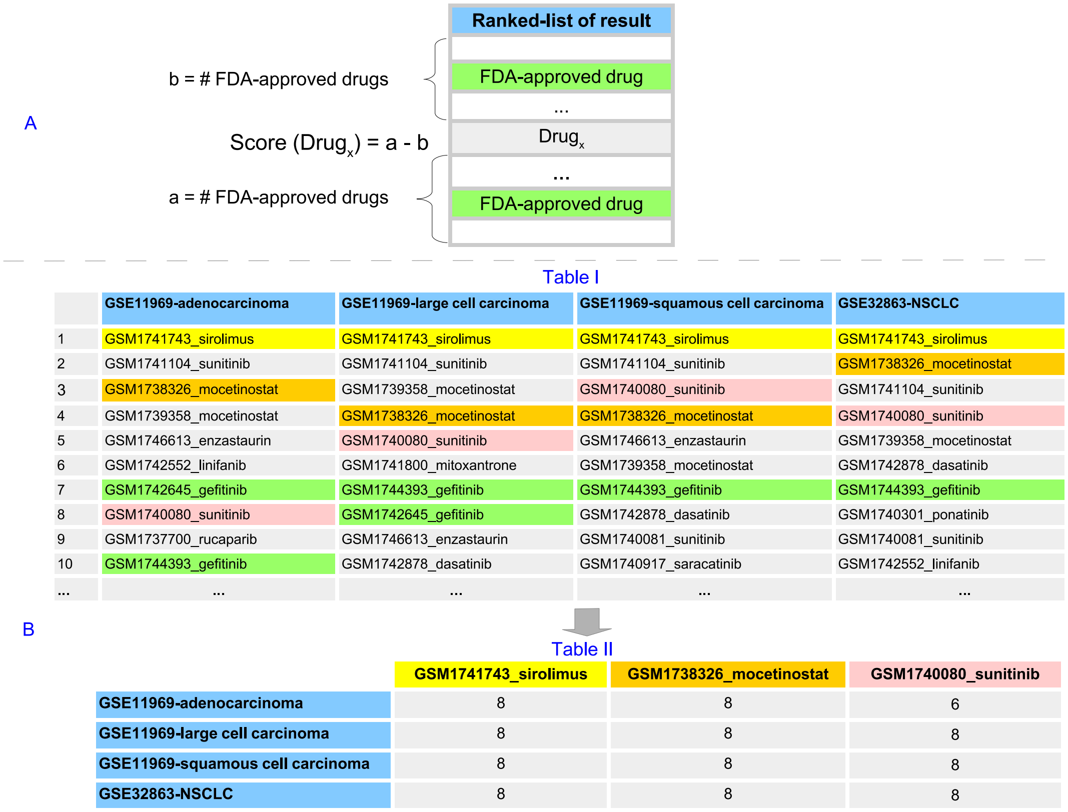


Table I

	GSE11969-adenocarcinoma	GSE11969-large cell carcinoma	GSE11969-squamous cell carcinoma	GSE32863-NSCLC
1	GSM1741743_sirolimus	GSM1741743_sirolimus	GSM1741743_sirolimus	GSM1741743_sirolimus
2	GSM1741104_sunitinib	GSM1741104_sunitinib	GSM1741104_sunitinib	GSM1738326_mocetinostat
3	GSM1738326_mocetinostat	GSM1739358_mocetinostat	GSM1740080_sunitinib	GSM1741104_sunitinib
4	GSM1739358_mocetinostat	GSM1738326_mocetinostat	GSM1738326_mocetinostat	GSM1740080_sunitinib
5	GSM1746613_enzastaurin	GSM1740080_sunitinib	GSM1746613_enzastaurin	GSM1739358_mocetinostat
6	GSM1742552_linifanib	GSM1741800_mitoxantrone	GSM1739358_mocetinostat	GSM1742878_dasatinib
7	GSM1742645_gefitinib	GSM1744393_gefitinib	GSM1744393_gefitinib	GSM1744393_gefitinib
8	GSM1740080_sunitinib	GSM1742645_gefitinib	GSM1742878_dasatinib	GSM1740301_ponatinib
9	GSM1737700_rucaparib	GSM1746613_enzastaurin	GSM1740081_sunitinib	GSM1740081_sunitinib
10	GSM1744393_gefitinib	GSM1742878_dasatinib	GSM1740917_saracatinib	GSM1742552_linifanib
...	...	...	...	...

Table II

	GSM1741743_sirolimus	GSM1738326_mocetinostat	GSM1740080_sunitinib
GSE11969-adenocarcinoma	8	8	6
GSE11969-large cell carcinoma	8	8	8
GSE11969-squamous cell carcinoma	8	8	8
GSE32863-NSCLC	8	8	8

Figure 3: Drug repurposing score computation. A) Given a ranked-list of drugs (drug instances) obtained by applying our approach on a disease dataset, a score is assigned to each drug indicating how better or worse that drug is ranked in comparison to already FDA-approved drugs. The score for $Drug_x$ is defined as $Score(Drug_x) = a - b$, where a and b denote the number of already FDA-approved drugs that are ranked worse and better than $Drug_x$, respectively. B) The non-small cell lung cancer (NSCLC): in total there are 8 drugs that are already FDA-approved for treatment of NSCLC. Table I shows the lists of 10 top-ranked drugs, results of the proposed approach using 4 NSCLC datasets: GSE11969-adenocarcinoma, GSE11969-large cell carcinoma, GSE11969-squamous cell carcinoma, and GSE32863-adenocarcinoma. The scores for GSM1741743_sirolimus, GSM1738326_mocetinostat, and GSM1740080_sunitinib across NSCLC datasets are summarized in Table II. A score of 8 means that each candidate ranked higher than all 8 instances of FDA-approved drugs.

Using this objective measure, we then calculate an average score for each drug across different disease datasets (Figure 2B). And finally, we compute an average score for each distinct drug across different instances, if there are multiple instances for that drug (Figure 2C). We select the top 5% drug candidates from the ranked lists obtained by applying our approach on disease datasets and rank such drugs based on the scores computed by the this method, from highest to the lowest.

2.5 Discussion and results

To validate our approach, we analyzed 19 datasets from four different conditions: idiopathic pulmonary fibrosis (IPF) (6 datasets), non-small cell lung cancer (NSCLC) (4 datasets), prostate cancer (3 datasets), and breast cancer (6 datasets).

We compare the results of 3 computational drug repurposing approaches: our system-level approach, the most popular approach proposed by [248] (henceforth drug-disease), and a classical method based on disease and drug signature anti-correlation (henceforth anti-correlation).

Both the drug-disease and the anti-correlation approaches are based on the hypothesis that if gene expression signature is perturbed in one direction in a disease state, and in the opposite (reverse) direction upon a drug exposure, then that drug may have the potential therapeutic effect for the disease. The difference between the two approaches is related on the approach used to calculated the match between a disease and a drug. Given a disease gene expression signature (query signature) and a drug gene expression signatures (reference signature), the Sirota et al.'s drug-disease similarity approach calculate an enrichment score for the up-regulated and down-regulated disease genes (by applying a Kolmogorov-Smirnov (KS) test). We use the R implementation of this approach available in the package DrugVsDisease [197].

In contrast, the classical anti-correlation method calculates a similarity score for drug-disease pairs by computing the Pearson correlation coefficient between the drug gene expression signature and the given disease gene expression signature. Drugs are ranked from

Table 2: Preliminary support by preclinical or clinical studies showing the therapeutic potential of the proposed candidates. These candidates are currently FDA-approved but for other indications.

Disease	Proposed candidate	Preclinical / clinical evidence	ClinicalTrials.gov ID
IPF	Sunitinib	[92, 143, 222]	
	Dabrafenib	[163, 193, 301]	
	Nilotinib	[1, 8, 15, 38, 46, 91, 104, 220]	
NSCLC	Sunitinib	[189, 251]	NCT00092001, NCT00372775, NCT00693992, NCT00864721
	Sirolimus	[30, 77, 90, 234]	NCT00923273
	Everolimus	[90, 212, 252]	NCT01061788
	Ponatinib	[44, 45, 81, 84, 218, 263]	NCT01813734
Prostate cancer	Podophyllotoxin	[25, 48, 53, 88, 109, 137, 159]	
	Acetylsalicylic acid	[29, 52, 97, 118, 160, 188, 229, 249]	NCT02757365,NCT03103152,NCT02804815
	Papaverine	[89, 112, 236]	
	Mefloquine	[85, 247, 294]	
	Vorinostat	[32, 39, 66, 130]	NCT00330161,NCT00589472
	Sirolimus	[9, 40, 114, 215]	NCT00311623,NCT02565901
Breast cancer	Captopril	[123, 129, 144, 185, 225, 230]	NCT00086723
	Glibenclimiade	[2, 190, 202, 213, 224, 299]	
	Fluorometholone	[127, 132, 155]	
	Etoposide	[18, 305, 306]	NCT00026949, NCT01492556, NCT01589159
	Colchicine	[255]	
	Tretinoin	[33, 82, 157, 199, 256]	

the highly anti-correlated to the highly correlated, according to their score.

The Anatomical Therapeutic Chemical ATC drug classification system, recommended by the World Health Organization (WHO) (http://www.whocc.no/atc/) provides a useful classification information based on drugs mechanism of actions. However, it is not precise enough to be considered as the gold standard. Drugs from the same class may not have the exact therapeutic desired effect for a condition. For instance, Nintedanib is an antineoplastic agent in L01 ATC class and it is the FDA-approved drug for treatment of IPF. However, not all drugs in the antineoplastic class have promising therapeutic effects on IPF because of pulmonary toxicities [69]. Hence, using the ATC class to repurpose drugs or even validate drug repurposing results is not feasible.

In this study, we compare the various approaches based on their ability to identify drugs that have already been FDA-approved for that condition (gold standard), based exclusively on the molecular data. In essence, a good repurposing approach should place already approved drugs at the very top of the list of drugs proposed for that particular disease. We used the Wilcoxon rank sum test [285] to determine whether the proposed approach is significantly better than the existing approaches.

Table 2 shows the proposed candidates for treatment of four human diseases: IPF, NSCLC, prostate cancer, breast cancer, and preliminary evidences that support the usefulness of those candidates in treatment of the given diseases.

2.5.1 Idiopathic pulmonary fibrosis

The list of IPF datasets we used in our analysis is summarized in Table 3. We compare the results of our approach with the existing approach proposed by [248] (drug-disease), as well as the classical method (anti-correlation). The lists of the top 10 drugs are summarized in Table 4.

Table 3: Idiopathic pulmonary fibrosis (IPF) datasets

Dataset	Source	Samples
GSE1724	NCBI GEO [219]	Treated (TGFbeta) vs untreated
GSE21369	NCBI GEO [51]	Interstitial lung disease vs healthy
GSE24206-advanced	NCBI GEO [176]	Advance stage IPF vs healthy
GSE24206-early	NCBI GEO [176]	Early stage IPF vs healthy
GSE44723	NCBI GEO [204]	Rapid progressing IPF vs healthy
LGRC	Lung Genomics Research Consortium	Interstitial lung disease vs healthy

Gold standard: Nintedanib is the only FDA-approved drug for IPF in our drug input datasets (highlighted as green). It inhibits *RTKs* such as *PDGFR* (α , β), *FGFR(1,2,3)*, *VEGFR(1,2,3)*, and *FLT*3, among them, *FGFR*, *PDGFR*, and *VEGFR* have been implicated in the pathogenesis of IPF. Additionally, Nintedanib inhibits *nRTKs* such as *Lck*, *Lyn* and *Src kinases* [91, 103, 131, 171, 214, 254, 289].

We select the top 5% of drugs ranked lists obtained by applying the proposed approach on 6 IPF datasets. As shown in Table 5, these drugs are ranked from the highest to the lowest, based on their scores. The score assigned to Nintedanib is 0. Since we have four instances for Nintedanib, drugs scores range between -4 and 4. The largest negative score

Table 4: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using IPF datasets (the top 10 drugs). The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach and drug-disease approach using datasets GSE24206-early, GSE24206-advanced, GSE44723, GSE21369, LGRC- ILD, GSE1724 datasets are 0.02, 0.02, 0.01, 0.02, 0.01, 0.01, respectively. Using these datasets. The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach and Anti-correlation approach using are 0.02, 0.02, 0.01, 0.02, 0.01, 0.01, respectively. Drugs highlighted with green are FDA-approved for the treatment of IPF. The * denotes the drugs that are currently FDA-approved but for other indications. The proposed approach was the only one that was able to rank the FDA-approved Nintedanib in the top 10. In contrast, none of the existing approaches was able to retrieve the FDA-approved drug in any of these 6 datasets.

Proposed	Drug-disease	Anti-correlation	Proposed	Drug-disease	Anti-correlation
GSE24206-early			GSE24206-advanced		
GSM1740570_saracatinib	GSM1746916_radicicol	GSM1746916_radicicol	GSM1740570_saracatinib	GSM1746916_radicicol	GSM1746916_radicicol
GSM1743214_nintedanib	GSM1746864_radicicol	GSM1746864_radicicol	GSM1743214_nintedanib	GSM1746864_radicicol	GSM1746864_radicicol
GSM1742836_celastrol	GSM1738326_mocetinostat*	GSM1746893_radicicol	GSM1745714_buparlisib	GSM1738291_azacitidine*	GSM1746893_radicicol
GSM1745714_buparlisib	GSM1738291_azacitidine*	GSM1742836_celastrol	GSM1742836_celastrol	GSM1746893_radicicol	GSM1738290_azacitidine*
GSM1742552_linifanib	GSM1738794_garcinol*	GSM1738290_azacitidine*	GSM1742552_linifanib	GSM1738772_ischemin	GSM1742836_celastrol
GSM1742850_nintedanib	GSM1745213_nilotinib*	GSM1738291_azacitidine*	GSM1740917_saracatinib	GSM1742836_celastrol	GSM1738291_azacitidine*
GSM1740917_saracatinib	GSM1737397_quizartinib*	GSM1745530_nilotinib*	GSM1739549_saracatinib	GSM1737397_quizartinib*	GSM1742552_linifanib
GSM1740731_CH5424802	GSM1743996_sirolimus*	GSM1745213_nilotinib*	GSM1742850_nintedanib	GSM1738326_mocetinostat*	GSM1743996_sirolimus*
GSM1743268_linifanib	GSM1742836_celastrol	GSM1742552_linifanib	GSM1740731_CH5424802	GSM1742718_sorafenib*	GSM1742716_sorafenib*
GSM1739549_saracatinib	GSM1746893_radicicol	GSM1742716_sorafenib*	GSM1745213_nilotinib*	GSM1746811_ruxolitinib*	GSM1745530_nilotinib*
GSE44723			GSE21369		
GSM1741104_sunitinib*	GSM1737411_NVP-BGT226	GSM1737409_NVP-BGT226	GSM1737700_rucaparib*	GSM1737352_everolimus*	GSM1737700_rucaparib*
GSM1740080_sunitinib*	GSM1743823_fostamatinib*	GSM1737411_NVP-BGT226	GSM1740570_saracatinib	GSM1737699_rucaparib*	GSM1737699_rucaparib*
GSM1742552_linifanib	GSM1745509_NVP-BEZ235	GSM1740923_BI-2536*	GSM1745714_buparlisib	GSM1737700_rucaparib*	GSM1738767_decitabine*
GSM1744393_gefitinib*	GSM1737409_NVP-BGT226	GSM1740576_BI-2536*	GSM1740731_CH5424802	GSM1738767_decitabine*	GSM1740731_CH5424802
GSM1737353_everolimus*	GSM1738100_tranylcypromine*	GSM1745509_NVP-BEZ235	GSM1745213_nilotinib*	GSM1746916_radicicol	GSM1737385_motesanib*
GSM1742436_GDC-0941	GSM1741779_vorinostat*	GSM1740570_saracatinib	GSM1742504_nintedanib	GSM1737385_motesanib*	GSM1737624_entinostat
GSM1743268_linifanib	GSM1742856_canertinib*	GSM1737412_NVP-BGT226	GSM1737448_idelalisib*	GSM1742800_palbociclib*	GSM1744048_imatinib*
GSM1743214_nintedanib	GSM1742795_palbociclib*	GSM1737353_everolimus*	GSM1745530_nilotinib*	GSM1737990_mocetinostat*	GSM1737698_rucaparib*
GSM1741743_sirolimus*	GSM1739241_olaparib*	GSM1745194_NVP-BEZ235	GSM1742836_celastrol	GSM1737443_idelalisib*	GSM1738290_azacitidine*
GSM1740570_saracatinib	GSM1740387_CH5424802	GSM1738308_entinostat	GSM1739679_dabrafenib*	GSM1742836_celastrol	GSM1741754_sirolimus*
LGRC-ILD			GSE1724		
GSM1740570_saracatinib	GSM1738326_mocetinostat*	GSM1738326_mocetinostat*	GSM1741104_sunitinib*	GSM1746800_nilotinib*	GSM1746916_radicicol
GSM1739549_saracatinib	GSM1742795_palbociclib*	GSM1737624_entinostat	GSM1740570_saracatinib	GSM1746916_radicicol	GSM1742552_linifanib
GSM1740917_saracatinib	GSM1737410_NVP-BGT226	GSM1739358_mocetinostat*	GSM1739549_saracatinib	GSM1743953_linifanib	GSM1742836_celastrol
GSM1743214_nintedanib	GSM1741779_vorinostat*	GSM1739435_belinostat*	GSM1742552_linifanib	GSM1742836_celastrol	GSM1745213_nilotinib*
GSM1743268_linifanib	GSM1737624_entinostat	GSM1741779_vorinostat*	GSM1745714_buparlisib	GSM1745958_dasatinib*	GSM1743268_linifanib
GSM1742552_linifanib	GSM1737411_NVP-BGT226	GSM1737410_NVP-BGT226	GSM1743268_linifanib	GSM1743268_linifanib	GSM1746864_radicicol
GSM1742504_nintedanib	GSM1739435_belinostat*	GSM1746864_radicicol	GSM1740917_saracatinib	GSM1741215_veliparib*	GSM1745626_vemurafenib
GSM1744170_GDC-0941	GSM1737409_NVP-BGT226	GSM1741767_vorinostat*	GSM1740080_sunitinib*	GSM1741184_regorafenib*	GSM1743197_celastrol
GSM1745714_buparlisib	GSM1737455_vandetanib*	GSM1737409_NVP-BGT226	GSM1742706_alvocidib	GSM1747067_mitoxantrone*	GSM1744371_selumetinib
GSM1741265_saracatinib	GSM1738350_pracinostat	GSM1737642_mocetinostat*	GSM1742850_nintedanib	GSM1741566_veliparib*	GSM1746881_mitoxantrone*

Table 5: The top 5% drugs obtained from the result of our repurposing approach. These drugs are ranked based on the scores generated by the systematic method. The * denotes the drugs that are currently FDA-approved but for other indications. The score for Nintedanib, the FDA-approved drug for IPF, is 0. Drugs with the same scores are sorted based on their average ranks. In this analysis, the scores computed by our systematic method can be further normalized based on dividing each score by the total number of FDA-approved drug instances for the disease of interest (e.g. 4 FDA-approved drug instances for IPF). Therefore, the scores calculated for each drug across different diseases will be in the same range (-1,1). For instance, the normalized score for Saracatinib will be 0.37.

Drug	Score
Saracatinib	1.5
Nintedanib	0
Linifanib	-0.67
Sunitinib *	-1.42
Buparlisib	-1.83
GDC-0941	-1.92
Alvocidib	-2.58
Dabrafenib *	-2.67
Nilotinib *	-2.83
Gefitinib *	-2.92
Idelalisib *	-2.92
CH5424802	-3
Everolimus *	-3
Dovitinib	-3
Rucaparib *	-3.08
Celastrol	-3.08
NVP-BEZ235	-3.17
Selumetinib	-3.17
Erlotinib *	-3.58
Sirolimus *	-3.58

for a drug indicates that drug ranked worse than all four instances of Nintedanib. Drugs with the same scores are sorted based on their average ranks in drugs ranked lists.

In this analysis, the scores computed by our systematic method can be further normalized based on dividing each score by the total number of FDA-approved drug instances for the disease of interest (e.g. 4 FDA-approved drug instances for IPF). Therefore, the scores calculated for each drug across different diseases will be in the same range (-1,1). For instance, the normalized score for Saracatinib will be 0.37.

Proposed candidates: We propose Sunitinib (p = 0.0009), Dabrafenib (p = 0.0009), and Nilotinib (p = 0.0009) as repurposing candidates for treatment of IPF. Saracatinib, Linifanib, Buparlisib, GDC-0941, and Alvocidib are also highly ranked by our approach for treatment of IPF. Although these drugs are not approved by FDA yet, they can be considered for further experimental tests.

Sunitinib is a small molecule that inhibits multiple receptor tyrosine kinases (RTKs), including vascular endothelial growth factor receptors (VEGFR) and platelet-derived growth factor receptors (PDGFR). It is approved by FDA for the treatment of Gastrointestinal stromal tumor, advanced renal cell carcinoma, and progressive well-differentiated pancreatic neuroendocrine tumors [64, 183]. It was investigated for its anti-fibrotic and anti-angiogenic properties. Its efficiency was experimentally proved in a bleomycin-induced mouse model and it has been proposed for the treatment of IPF [143]. Results of *in vitro* studies and animal models show that receptor tyrosine kinases, such as *PDGFR*, *VEGFR* and *FGFR*, and non-receptor tyrosine kinases, such as the *Src* family, play crucial roles in the pathogenesis of IPF [92, 222].

Dabrafenib is approved by FDA for the treatment of patients with unresectable or metastatic melanoma. Recent clinical studies demonstrate that the extracellular signal regulated kinase (ERK) and mitogen-activated protein kinase (MAPK) are up-regulated in lung tissues of patients with IPF [163, 301]. In particular, results of studies on MAPK signaling pathways show that the level of serine/threonine-protein kinase B-Raf (BRAF) is increased in patients samples compared to the normal ones, suggesting the potential therapeutic effects of MEK/ERK inhibitors for pulmonary fibrosis [163, 193]. This supports the idea that the BRAF inhibitor Dabrafenib may have atherapeutic effect on IPF.

Nilotinib is another FDA-approved drug we propose to be repurposed for the treatment of IPF. Nilotinib is a transduction inhibitor targeting *BCR-ABL*, *c-kit*, and *PDGF*, that is approved by FDA for treatment of patients who are newly diagnosed with Philadelphia chromosome positive chronic myeloid leukemia (Ph+CML). It is also used for treatment of patients with Ph+CML in chronic phase and accelerated phase if they were resistant (or intolerant) to previous treatments. The potential roles of *PDGFs* in IPF have been shown by many studies [8, 15, 38, 104, 289]. The advantage of *PDGF* inhibition in IPF is well studied and supported by several studies [1, 46, 281, 289]. Authors of [91, 220] confirmed the potential effect of Nilotinib in decreasing the extent of pulmonary fibrosis in a mouse model.

The phosphatidylinositol 3 kinase (PI3K) inhibitors **Buparlisib** and GDC-0941 are undergoing clinical trials for a number of diseases. Buparlisib is in Phase III of clinical trials for treatment of breast cancer and in and Phase II for several other solid tumors. GDC-0941(Pictilisib) has been used in clinical trials for the treatment of several cancers, including breast cancer. Preclinical studies proved that PI3K inhibitors have potential roles in treatment of IPF by interfering with the fibrogenic effects of $TGF-\beta1$ signaling [26, 59, 107, 178]. Based on this evidence, Buparlisib and GDC-0941 may have potential therapeutic effects on IPF.

The tyrosine kinase inhibitors **Saracatinib** and **Linifanib** are also highly ranked by our approach for treatment of IPF. Saracatinib (AZD0530) is an oral, tyrosine kinase inhibitor selective for *Src*. It underwent clinical tests at AstraZeneca for the treatment of cancer [94, 148, 179, 211]. However, it failed to show a sufficient efficacy in these studies. Subsequently, it was proposed for other usages such as Alzheimer's disease (in Phase

II) [191]. Linifanib (ABT-869) is also a multi-targeted receptor tyrosine kinase inhibitor that is intended to suppress tumor growth. It is investigated for treatment of leukemia (myeloid), myelodysplastic syndrome, and solid tumors [47, 50, 279]. The efficiency and tolerability of Linifanib versus Sorafenib has been assessed in patients with advanced hepatocellular carcinoma [36]. The tyrosine kinase inhibitors are proven to be effective in treatment of IPF [4, 6, 26, 91, 152, 222, 288]. In particular, the Src kinase inhibitor Saracatinib is reported to be useful in treatment of IPF through targeting the $TGF-\beta$ signaling pathway [108].

The Food and Drug Administration's Office of Orphan Products Development provides orphan drug status to medicines that are designed for the treatment, diagnosis or prevention of rare conditions or diseases that affect fewer than 200,000 people in the U.S or that affect more than 200,000 people but are not expected to recover the costs of developing and marketing a treatment drug [246, 273]. IPF has a significant affect on patients' lives and finding optimal treatments for this condition is extremely important (https://pulmonaryfibrosisnews.com). More specifically, Saracatinib is currently in (STOP IPF) trial for the treatment of patients with IPF where 100 participants with IPF will receive either Saracatinib or placebo for 24 weeks [6]. In this trial, the safety and tolerability of Saracatinib as well as the early indicators of Saracatinib efficacy and the relevant biomarkers of Src kinase activity and fibrogenesis in treatment of IPF will be evaluated [6].

Alvocidib is a cyclin-dependent kinase (CDK) inhibitor that is undergoing clinical trials for a number of cancers: esophageal cancer, leukemia, lung cancer, liver cancer, and lymphoma. Studies of murine models show that the CDK inhibitors block the epithelial apoptosis and decrease the tissue fibrosis in pulmonary fibrosis [115, 150]. As a result, CDK inhibitors have been suggested as a novel therapeutic strategy against IPF [310].

Drug-disease networks

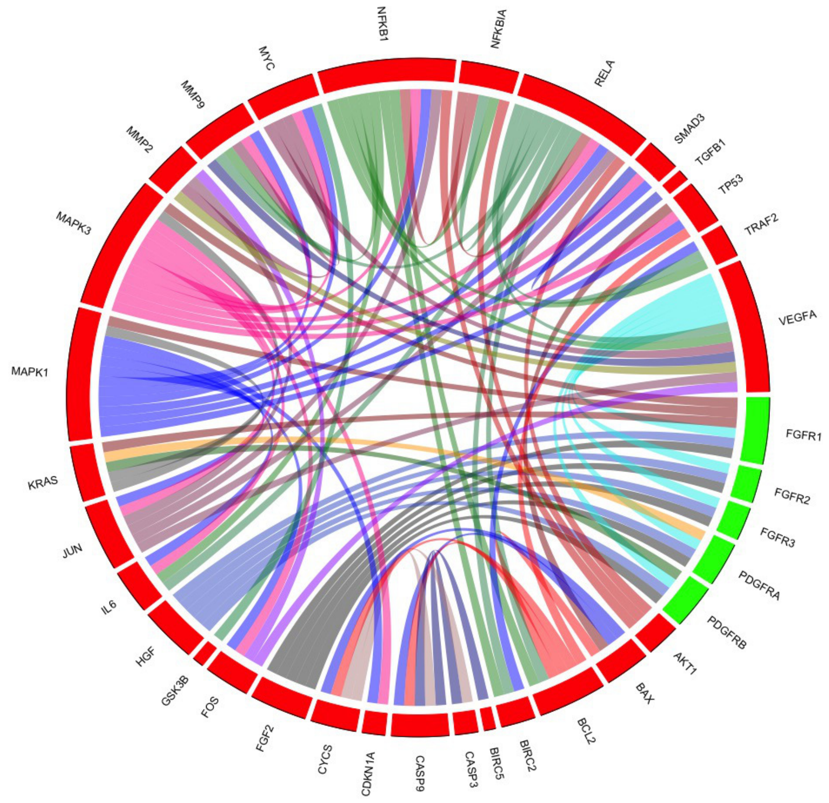


Figure 4: The chord diagram represents the subnetwork of DDN for Nintedanib, the FDA-approved drug for IPF. In order to obtain this subnetwork, we used IPF-associated genes that are included in KEGG's *Pathways in cancer* (the target pathway for Nintedanib). Sectors and chords represent the genes and associations between the genes in the network, respectively. Red sectors represent genes known to be associated to IPF disease. Sectors representing the Nintedanib target genes are green.

We used chord diagrams to represent subnetworks of drug-disease networks (DDN) for Nintedanib (Figure 4), Sunitinib (Figure 5), Linifanib (Figure 6), and Saracatinib (Figure 7). In order to obtain the subnetworks we use IPF-associated genes belonging to the *Pathways in cancer*, the target pathway for Nintedanib. The subnetwork S= (V,E) with the node set V and edge set E is represented as follow:

$$S = (V, E) : (V \subset Path \cap (Disease_t \cup Drug_t)) \wedge (E \subset DDN) \tag{2.3}$$

where $Disease_t = \{x_1, x_2, ..., x_n\}$, $Drug_t = \{y_1, y_2, ..., y_n\}$, and $Path$ denote the disease-related genes, drug targets, and the genes on the *Pathways in cancer*, respectively.

In the chord diagram, sectors represent the genes and the chords represent the associations between various genes in the network we built. The red sectors represent the genes known to be associated to IPF. The green sectors represent the genes targeted by Nintedanib.

In addition to chord diagrams, we used the edge lists to represent the DDN subnetworks we construct for the repurposing candidates. In this list, the association between the genes is represented as a tuple (e.g. FGFR1 - KRAS). Table 6 shows the edge lists representing DDN subnetworks for repurposing drugs: Nintedanib, Nilotinib, and Sunitinib.

Table 7 shows the edge lists representing DDN subnetworks for drugs: Linifanib and Saracatinib. These drugs are currently undergoing clinical trials for several indications (see detail in section 3.1 of the manuscript). Red entires represent genes known to be associated to IPF disease. Entries representing the Nintedanib target genes are green.

Drug-drug networks

Figures 9 and 8 show the drug-drug networks we generated using the known knowledge (target pathways and target genes) of Nintedanib (FDA-approved for IPF treatment) and top-ranked candidates for IPF, where circles correspond to drugs, and two drugs being connected if they share target pathways or target genes, respectively. The target path-

Table 6: The edge lists represent the subnetwork of DDN of drugs: Nintedanib, Nilotinib, and Sunitinib. Nintedanib is the FDA-approved drug for IPF treatment. Nilotinib and Sunitinib are the repurposing candidates for the treatment of IPF. In this list, the association between two genes is represented as gene pairs. In order to obtain the subnetworks, we used IPF-associated genes that are included in KEGG's *Pathways in cancer* (the target pathway for Nintedanib). Red entires represent genes known to be associated to IPF disease. Entries representing the Nintedanib target genes are green.

Nintedanib				Nilotinib				Sunitinib			
Gene1 → Gene2		Gene1 → Gene2		Gene1 → Gene2		Gene1 → Gene2		Gene1 → Gene2		Gene1 → Gene2	
FGFR1	KRAS	JUN	VEGFA	PDGFRα	KRAS	MYC	VEGFA	FGFR1	KRAS	JUN	VEGFA
FGFR1	MAPK1	KRAS	MAPK1	PDGFRβ	KRAS	NFKB1	BCL2	FGFR1	MAPK1	KRAS	MAPK1
FGFR1	MAPK3	KRAS	MAPK3	FGF2	PDGFRα	NFKB1	BIRC2	FGFR1	MAPK3	KRAS	MAPK3
PDGFRα	KRAS	MAPK1	CDKN1A	FGF2	PDGFRβ	NFKB1	IL6	PDGFRα	KRAS	MAPK1	CDKN1A
PDGFRβ	KRAS	MAPK1	IL6	HGF	PDGFRα	NFKB1	MMP9	PDGFRβ	KRAS	MAPK1	IL6
FGF2	FGFR1	MAPK1	MMP9	HGF	PDGFRβ	NFKB1	NFKBIA	FGF2	PDGFRα	MAPK1	MMP9
FGF2	FGFR2	MAPK1	MYC	VEGFA	PDGFRα	NFKB1	TRAF2	FGF2	PDGFRβ	MAPK1	MYC
FGF2	FGFR3	MAPK1	NFKB1	VEGFA	PDGFRβ	NFKB1	VEGFA	HGF	FGFR1	MAPK1	NFKB1
FGF2	PDGFRα	MAPK1	RELA	AKT1	NFKB1	RELA	BCL2	HGF	FGFR2	MAPK1	RELA
FGF2	PDGFRβ	MAPK1	TP53	AKT1	NFKBIA	RELA	BIRC2	HGF	PDGFRα	MAPK1	TP53
HGF	FGFR1	MAPK3	CDKN1A	AKT1	RELA	RELA	IL6	HGF	PDGFRβ	MAPK3	CDKN1A
HGF	FGFR2	MAPK3	IL6	BAX	CASP9	RELA	MMP9	VEGFA	FGFR1	MAPK3	IL6
HGF	FGFR3	MAPK3	MMP9	BAX	CYCS	RELA	NFKBIA	VEGFA	FGFR2	MAPK3	MMP9
HGF	PDGFRα	MAPK3	MYC	BIRC5	CASP9	RELA	TRAF2	VEGFA	FGFR3	MAPK3	MYC
HGF	PDGFRβ	MAPK3	NFKB1	CASP9	CASP3	RELA	VEGFA	VEGFA	PDGFRα	MAPK3	NFKB1
VEGFA	FGFR1	MAPK3	RELA	CYCS	CASP3	TGFB1	SMAD3	VEGFA	PDGFRβ	MAPK3	RELA
VEGFA	FGFR2	MAPK3	TP53	FOS	VEGFA			FGF2	FGFR1	MAPK3	TP53
VEGFA	FGFR3	MYC	MMP2	JUN	VEGFA			FGF2	FGFR2	MYC	MMP2
VEGFA	PDGFRα	MYC	MMP9	KRAS	MAPK1			AKT1	BCL2	MYC	MMP9
VEGFA	PDGFRβ	MYC	VEGFA	KRAS	MAPK3			AKT1	NFKB1	MYC	VEGFA
AKT1	NFKB1	NFKB1	BCL2	MAPK1	CDKN1A			AKT1	NFKBIA	NFKB1	BCL2
AKT1	NFKBIA	NFKB1	BIRC2	MAPK1	IL6			AKT1	RELA	NFKB1	BIRC2
AKT1	RELA	NFKB1	IL6	MAPK1	MMP9			BAX	CASP9	NFKB1	IL6
BAX	CASP9	NFKB1	MMP9	MAPK1	MYC			BAX	CYCS	NFKB1	MMP9
BAX	CYCS	NFKB1	NFKBIA	MAPK1	NFKB1			BIRC5	CASP9	NFKB1	NFKBIA
BIRC5	CASP9	NFKB1	TRAF2	MAPK1	RELA			CASP9	CASP3	NFKB1	TRAF2
CASP9	CASP3	NFKB1	VEGFA	MAPK1	TP53			CYCS	CASP3	NFKB1	VEGFA
CYCS	CASP3	RELA	BCL2	MAPK3	CDKN1A			FOS	IL6	RELA	BCL2
CYCS	CASP9	RELA	BIRC2	MAPK3	IL6			FOS	MMP2	RELA	BIRC2
FOS	MMP2	RELA	IL6	MAPK3	MMP9			FOS	MMP9	RELA	IL6
FOS	VEGFA	RELA	MMP9	MAPK3	MYC			FOS	VEGFA	RELA	MMP9
JUN	MMP2	RELA	NFKBIA	MAPK3	NFKB1			JUN	IL6	RELA	NFKBIA
JUN	NFKB1	RELA	TRAF2	MAPK3	RELA			JUN	MMP2	RELA	TRAF2
JUN	RELA	RELA	VEGFA	MAPK3	TP53			JUN	MMP9	RELA	VEGFA
JUN	VEGFA	TGFB1	SMAD3	MYC	MMP2			JUN	NFKB1	SMAD3	CDKN1A
KRAS	MAPK1	TRAF2	BIRC2	MYC	MMP9			JUN	RELA	TGFB1	SMAD3
										TP53	CDKN1A
										TRAF2	BIRC2
										TRAF2	NFKBIA

Table 7: The edge lists represent the subnetwork of DDN of drugs: Linifanib and Saracatinib. These drugs are currently undergoing clinical trials (for conditions other than IPF) that can be considered for the treatment of IPF. In this list, the association between two genes is represented as a tuple. In order to obtain the subnetworks, we used IPF-associated genes that are included in KEGG's *Pathways in cancer* (the target pathway for Nintedanib). Red entires represent genes known to be associated to IPF disease. Entries representing the Nintedanib target genes are green.

Linifanib	
Gene1 →	Gene2
PDGFRβ	KRAS
FGF2	PDGFRβ
HGF	PDGFRβ
VEGFA	PDGFRβ
AKT1	NFKB1
AKT1	NFKBIA
AKT1	RELA
BAX	CASP9
BAX	CYCS
BIRC5	CASP9
CASP9	CASP3
CYCS	CASP3
JUN	VEGFA
KRAS	MAPK1
KRAS	MAPK3
MAPK1	CDKN1A
MAPK1	IL6
MAPK1	MMP9
MAPK1	MYC
MAPK1	NFKB1
MAPK1	RELA
MAPK1	TP53
MAPK3	CDKN1A
MAPK3	IL6
MAPK3	MMP9
MAPK3	MYC
MAPK3	NFKB1
MAPK3	RELA
MAPK3	TP53
MYC	MMP2
MYC	MMP9
MYC	VEGFA
NFKB1	BCL2
NFKB1	BIRC2
NFKB1	NFKBIA
NFKB1	TRAF2
NFKB1	VEGFA
RELA	BCL2
RELA	BIRC2
RELA	NFKBIA
RELA	TRAF2
RELA	VEGFA
TGFB1	SMAD3

Saracatinib	
Gene1 →	Gene2
FGFR1	MAPK1
FGFR1	MAPK3
PDGFRα	KRAS
FGF2	PDGFRα
FGF2	PDGFRβ
HGF	PDGFRα
HGF	PDGFRβ
VEGFA	FGFR1
AKT1	NFKB1
AKT1	NFKBIA
AKT1	RELA
BAX	CASP9
BAX	CYCS
BIRC5	CASP9
CASP9	CASP3
CYCS	CASP3
FGF2	FGFR1
FOS	IL6
FOS	VEGFA
HGF	FGFR1
JUN	NFKB1
JUN	RELA
JUN	VEGFA
KRAS	MAPK1
KRAS	MAPK3
MAPK1	CDKN1A
MAPK1	IL6
MAPK1	MMP9
MAPK1	MYC
MAPK1	NFKB1
MAPK1	RELA
MAPK1	TP53
MAPK3	CDKN1A
MAPK3	IL6
MAPK3	MMP9
MAPK3	MYC
MAPK3	RELA
MAPK3	TP53
MYC	MMP2
MYC	MMP9
MYC	VEGFA
NFKB1	BCL2
NFKB1	BIRC2
NFKB1	BIRC2
NFKB1	IL6
NFKB1	MMP9
NFKB1	NFKBIA
NFKB1	TRAF2
NFKB1	VEGFA
RELA	BCL2
RELA	BIRC2
RELA	IL6
RELA	MMP9
RELA	NFKBIA
RELA	TRAF2
RELA	VEGFA
TGFB1	SMAD3

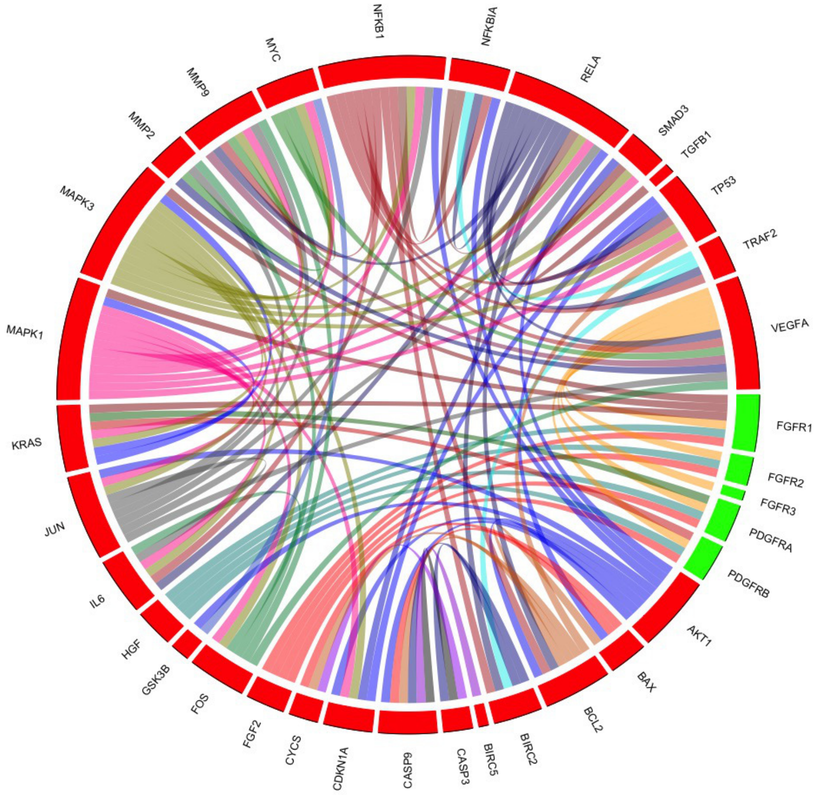


Figure 5: The chord diagram represents the subnetwork of DDN for Sunitinib, the repurposing candidate for the treatment of IPF. In order to obtain this subnetwork, we used IPF-associated genes that are included in KEGG's *Pathways in cancer* (the target pathway for Nintedanib). Sectors and chords represent the genes and associations between the genes in the network, respectively. Red sectors represent genes known to be associated to IPF disease. Sectors representing the Nintedanib target genes are green.

ways and target genes of drugs (represented by rectangles) are obtained from KEGG and Drugbank [287].

2.5.2 Non-small cell lung cancer

We obtained four non-small cell lung cancer (NSCLC) datasets from Gene Expression Omnibus (GEO): GSE32863 (adenocarcinoma) [241] and GSE11969 (subtypes: adenocarcinoma, large cell carcinoma, squamous cell carcinoma) [258]. Adenocarcinoma (40%

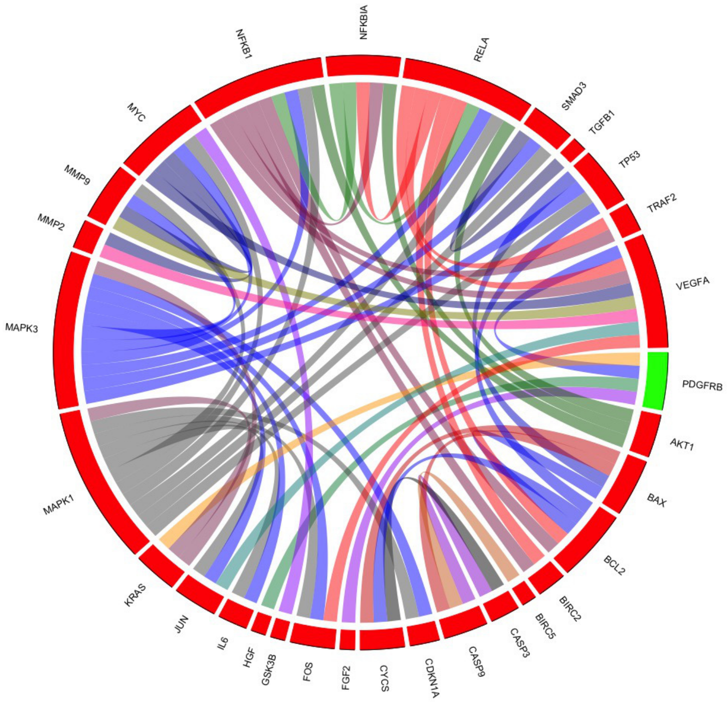


Figure 6: The chord diagram represents the subnetwork of DDN for Linifanib. In order to obtain this subnetwork, we used IPF-associated genes that are included in KEGG's *Pathways in cancer* (the target pathway for Nintedanib). Sectors and chords represent the genes and associations between the genes in the network, respectively. Red sectors represent genes known to be associated to IPF disease. Sectors representing the Nintedanib target genes are green.

of lung cancers), squamous cell carcinoma (25% of lung cancers), and large cell carcinoma (10% of lung cancers) are three main subtypes of NSCLC.

As shown in Panel A of Figure 3, given a ranked-list of drugs (drug instances) obtained by applying our approach on a disease dataset, a score for drug $Drug_x$ is defined as $Score(Drug_x) = a - b$, where terms a and b denote the number of already FDA-approved drugs (gold standards) that are ranked worse and better than $Drug_x$, respectively. Panel B in this figure shows the computed scores for a number of top-ranked dugs from non-small cell lung cancer (NSCLC) results. In total, there are 8 FDA-approved drugs for

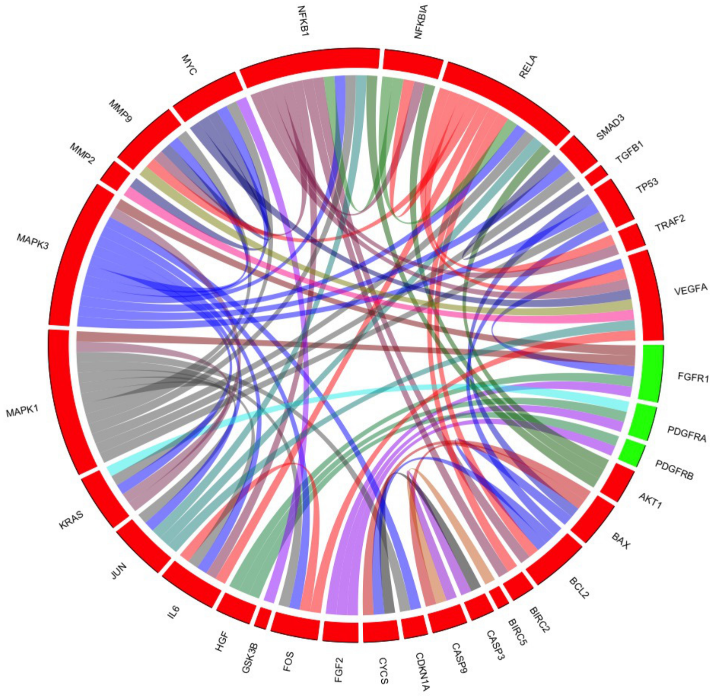


Figure 7: The chord diagram represents the subnetwork of DDN for Saracatinib. In order to obtain this subnetwork, we used IPF-associated genes that are included in KEGG's *Pathways in cancer* (the target pathway for Nintedanib). Sectors and chords represent the genes and associations between the genes in the network, respectively. Red sectors represent genes known to be associated to IPF disease. Sectors representing the Nintedanib target genes are green.

NSCLC in our drug input list. For instance, the score assigned to GSM1740080_sunitinib for the GSE11969-adenocarcinoma dataset is 6, as there is only one FDA-approved drug (highlighted with green) ranked better than GSM1740080_sunitinib and there are 7 FDA-approved drugs ranked below it (7-1=6). A score of 8 means that each candidate ranked higher than all 8 instances of FDA-approved drugs. Table II in Panel B summarizes the scores assigned to three drug instances: GSM1741743_sirolimus, GSM1738326_mocetinostat, and GSM1740080_sunitinib using four NSCLC datasets. We then calculate an average score for each drug across different disease datasets, and different instances, if there are

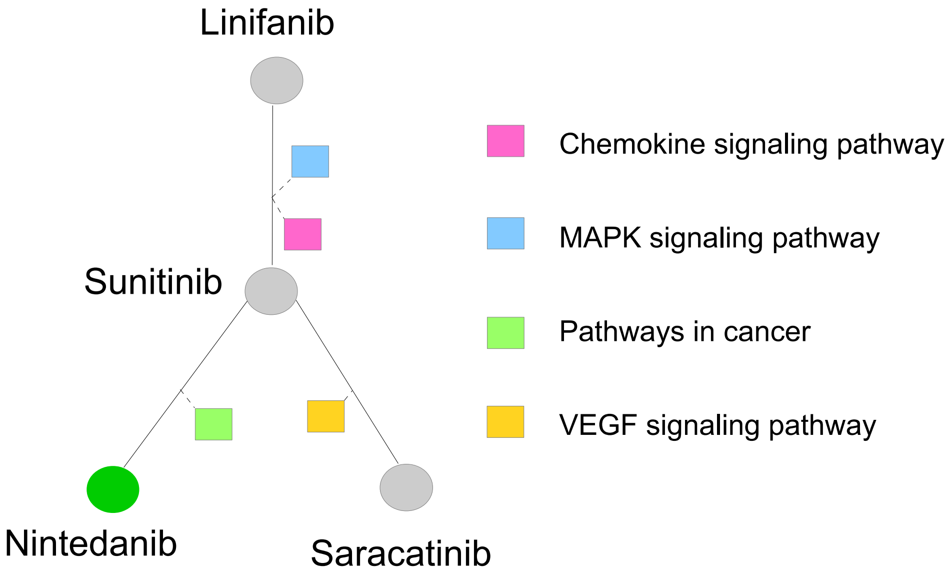


Figure 8: Drug-drug network is generated using the known knowledge of drugs target pathways, obtained from KEGG. Circles and rectangles correspond to drugs and target pathways, respectively. Drugs are connected to each other based on their common target pathways. Nintedanib (shown with green circle) is FDA-approved for treatment of IPF. *Pathways in cancer* is known as the target pathway for Nintedanib.

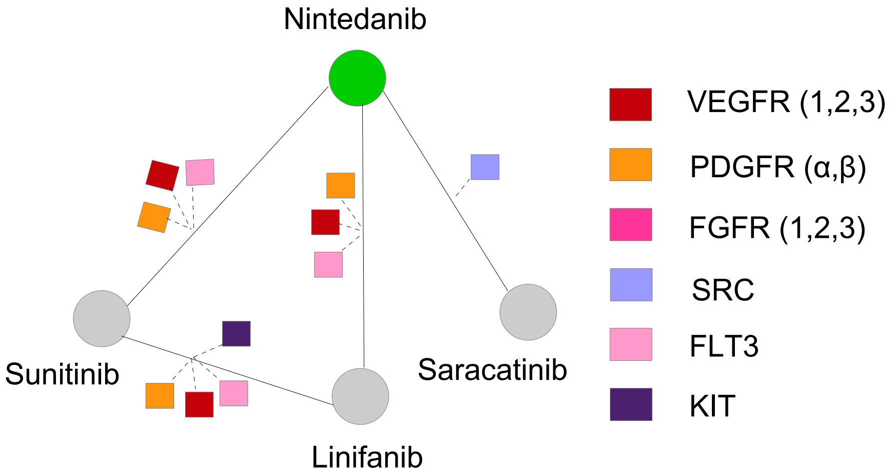


Figure 9: Drug-drug network is generated using the known knowledge of drugs target genes, obtained from KEGG and Drugbank [287]. Circles and rectangles correspond to drugs and target genes, respectively. Drugs are connected to each other based on their common target genes. Nintedanib (shown with green circle) is FDA-approved for treatment of IPF. *VEGFR(1,2,3)*, *PDGFR* (α,β), *FGFR (1,2,3)*, *SRC*, *FLT3*, and *KIT* are known to be target genes of Nintedanib [91, 103, 131, 171, 214, 254, 289].

multiple instances for that drug. This is shown in Figure 2.

We select the top 5% of drugs ranked lists obtained by applying our approach on disease datasets and rank such drugs based on the scores computed by the systematic method, from highest to the lowest.

Gold standards: Gefitinib and Crizotinib are the FDA-approved drugs for treatment of NSCLC. These drugs are included in our list of input drugs from LINCS. On all NSCLC datasets, three computational approaches were compared in terms of their ability to highly rank the instances of Gefitinib and Crizotinib that exist in the LINCS drug database.

Erlotinib is used as the maintenance treatment of locally advanced or metastatic NSCLC patients with no progress in their disease after four cycles of platinum-based first-line chemotherapy. It is also indicated after failure of at least one prior chemotherapy regimen in patients with NSCLC. Since Erlotinib is limited to very specific patients with NSCLC and none of our datasets fulfill such limitations, we did not consider it as the gold standard. However, the proposed approach is significantly better than the other two approaches even if Erlotinib were to be included as the gold standard.

Tables 9–12 show the results obtained on four NSCLC datasets using: i) the proposed approach, ii) the drug-disease approach, and iii) the anti-correlation approach. The FDA-approved drugs for NSCLC are highlighted in green.

We selected the top 5% of drugs ranked lists obtained by applying the proposed approach using 4 NSCLC datasets. As shown in Table 8, these drugs are ranked according to the scores computed by the systematic method from the highest to the lowest. The sum of the scores assigned to Gefitinib and Crizotinib (the FDA-approved drugs for NSCLC) is 0. Drugs with the same scores are sorted based on their average ranks in drugs ranked lists.

In this analysis, the scores computed by our systematic method can be further normalized based on dividing each score by the total number of FDA-approved drug instances for the disease of interest (8 FDA-approved drug instances for NSCLC). Therefore, the scores calculated for each drug across different diseases will be in the same range (-1,1).

Table 8: The top 5% of drugs obtained from the result of the proposed approach. These drugs are ranked based on the scores generated by the systematic method. The * denotes the drugs that are currently FDA-approved but for other indications. Gefitinib and Crizotinib, as highlighted with green, are FDA-approved for the treatment of NSCLC. Drugs with the same scores are sorted based on their average ranks in drugs ranked lists. In this analysis, the scores computed by our systematic method can be further normalized based on dividing each score by the total number of FDA-approved drug instances for the disease of interest (8 FDA-approved drug instances for NSCLC). Therefore, the scores calculated for each drug across different diseases will be in the same range (-1,1).

Drug	Score
Sunitinib *	4.625
Mocetinostat	2.75
Gefitinib	1.75
Roscovitine	-1
Sirolimus *	-1
Enzastaurin *	-1.75
Crizotinib	-1.75
Everolimus *	-2
Ponatinib *	-2.25
Saracatinib	-2.25
Rucaparib *	-3.125
Dasatinib *	-3.375
Linifanib	-3.625
Mitoxantrone *	-4.625

Proposed candidates: We propose the FDA-approved drugs: Sunitinib (p = 0.0009), Sirolimus (p = 0.0009), Enzastaurin (p = 0.0009), Everolimus (p = 0.001), and Ponatinib (p = 0.004) as repurposing candidates for treatment of NSCLC. Although Mocetinostat, Roscovitine, and Saracatinib are not approved by FDA yet, they can be prioritized for further investigations.

As discussed earlier, **Sunitinib** is an oral, small-molecule that inhibits RTKs, including VEGFR and PDGFR. Recent clinical trials have reported that Sunitinib has the provocative single-agent activity in previously treated patients with recurrent and advanced NSCLC [189, 251]. Sunitinib has completed the phase II of clinical trials for treatment of patients with NSCLC (ClinicalTrials.gov IDs: NCT00372775, NCT00092001, NCT00864721). The phase III of clinical trials on Sunitinib as a potential maintenance therapy in NSCLC patients has been completed. These patients had received four cycles of platinum-based chemotherapy without disease progression. The result of the trials has not published yet (ClinicalTrials.gov ID: NCT00693992).

Sirolimus, also known as Rapamycin, is a potent immunosuppressant that inhibits mammalian target of rapamycin (mTOR). The positive effect of Sirolimus in inhibiting the growth and progression of NSCLC is supported by several clinical studies [30, 77, 90, 234]. The phase I/II clinical trials have been launched to test the efficiency of Sirolimus in combination with Pemetrexed for treating patients with NSCLC (ClinicalTrials.gov ID: NCT01061788). The phase I clinical trials of Sunitinib and Sirolimus confirm that this combination is well-tolerated and warrants further investigation in advanced NSCLC (ClinicalTrials.gov ID: NCT00555256) [282].

Everolimus is a derivative of rapamycin (Sirolimus). It is approved by FDA for treatment of several conditions, including breast cancer, advanced renal cell carcinoma, renal angiomyolipoma, and tuberous sclerosis. It has shown antitumor activity both as the single agent and in combination with other agents in treatment of patients with NSCLC. Several clinical trials support the efficacy of Everolimus in treatment of NSCLC [90, 212, 252, 271] (ClinicalTrials.gov ID: NCT00096486).

Increased levels of protein kinase C (PKC) and AKT are known to be associated with the poor prognosis in NSCLC [57, 192]. **Enzastaurin**, an oral serine/threonine kinase inhibitor, suppresses PKC and protein kinase B/AK transforming (AKT) signaling, induces tumor cell apoptosis, and inhibits the proliferation and angiogenesis [192]. Enzastaurin is proven to inhibit the growth of NSCLC cell lines [184, 192, 261]. The Phase II evaluation of Enzastaurin as the second-and third- line treatment for NSCLC has completed with promising results (ClinicalTrials.gov ID: NCT00105092).

Fibroblast growth factor receptors (FGFRs) are known to be overexpressed in NSCLC [23, 235]. **Ponatinib** is proven to be effective against the FGFR1 kinase in 8p11 myeloproliferative syndrome (EMS) [45]. In particular, It has been shown that Ponatinib can suppress cell growth in NSCLC cell lines [218, 263]. Several studies reported that RET fusions are viable targets in NSCLC [44, 81, 84]. The phase II of the clinical trial (ClinicalTrials.gov ID: NCT01813734) is currently evaluating the safety and the effectiveness of the RET inhibitor Ponatinib in treating patients with NSCLC.

It has been reported that HDAC inhibitor **Mocetinostat** may restore normal cell function and reduce or inhibit the tumor growth [187]. A phase $1/2$ clinical trial of Mocetinostat, in combination with Durvalumab is currently ongoing in treating patients with solid tumors and NSCLC to evaluate the safety and efficacy of this combination [98] (ClinicalTrials.gov IDs: NCT02805660). Another clinical trial is currently undergoing to evaluate the clinical activity of Nivolumab in combination with three separate drugs, Glesatinib, Sitravatinib, or Mocetinostat in NSCLC (ClinicalTrials.gov ID: NCT02954991, phase II).

Roscovitine is an experimental drug in the class of pharmacological cyclin-dependent kinase (CDK) inhibitors. Current clinical studies [56, 195] suggest the combination of Roscovitine and Belinostat in treating patients with NSCLC. The phase II study of Roscovitine as a single agent in previously-treated patients with non-small cell lung cancer has

Table 9: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE11969-adenocarcinoma dataset (the top 10 drugs). The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach with drug-disease and anti-correlation approaches are 0.01 and 0.005, respectively. Drugs highlighted with green are FDA-approved for the treatment of NSCLC. The * denotes the drugs that are currently FDA-approved but for other indications.

GSE11969-adenocarcinoma		
Proposed	Drug-disease	Anti-correlation
GSM1741743_sirolimus*	GSM1746780_erlotinib*	GSM1737411_NVP-BGT226
GSM1741104_sunitinib*	GSM1738326_mocetinostat	GSM1739358_mocetinostat
GSM1738326_mocetinostat	GSM1739358_mocetinostat	GSM1741104_sunitinib*
GSM1739358_mocetinostat	GSM1741104_sunitinib*	GSM1740576_BI-2536
GSM1746613_enzastaurin	GSM1745191_NVP-BEZ235	GSM1738326_mocetinostat
GSM1742552_linifanib	GSM1745674_dovitinib	GSM1740923_BI-2536
GSM1742645_gefitinib	GSM1742797_palbociclib*	GSM1737409_NVP-BGT226
GSM1740080_sunitinib*	GSM1745194_NVP-BEZ235	GSM1742797_palbociclib*
GSM1737700_rucaparib*	GSM1746864_radicicol	GSM1737349_everolimus*
GSM1744393_gefitinib	GSM1741767_vorinostat*	GSM1745194_NVP-BEZ235

terminated with no data reported (ClinicalTrials.gov ID: NCT00372073).

Saracatinib is an inhibitor of SRC kinases that may improve NSCLC treatment [227, 228, 309]. It is undergoing phase II of clinical trials in treatment of patients with NSCLC (NCT00638937).

2.5.3 Prostate cancer

We use three different prostate cancer datasets. The first dataset is obtained by comparing gene expression levels between prostate tissues from 6 prostate cancer samples with 6 healthy samples using Affymetrix Human Genome U133 Plus2.0 Array. This dataset is available via GEO (GSE26910) [209]. GSE6919 is the second dataset that compares 65 primary prostate cancer samples with 18 healthy samples using Affymetrix Human Genome U95A Version 2 Array [43, 303]. The third dataset is the result of comparing gene expres-

Table 10: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE11969-large cell carcinoma dataset (the top 10 drugs). The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach with drug-disease and anti-correlation approaches are 0.001 and 0.0005, respectively. Drugs highlighted with green are FDA-approved for the treatment of NSCLC. The * denotes the drugs that are currently FDA-approved but for other indications.

GSE11969-large cell carcinoma		
Proposed	Drug-disease	Anti-correlation
GSM1741743_sirolimus*	GSM1745191_NVP-BEZ235	GSM1737411_NVP-BGT226
GSM1741104_sunitinib*	GSM1738326_mocetinostat	GSM1740576_BI-2536
GSM1739358_mocetinostat	GSM1739358_mocetinostat	GSM1740923_BI-2536
GSM1738326_mocetinostat	GSM1745194_NVP-BEZ235	GSM1745191_NVP-BEZ235
GSM1740080_sunitinib*	GSM1745149_GDC-0980	GSM1745194_NVP-BEZ235
GSM1741800_mitoxantrone*	GSM1741767_vorinostat*	GSM1739358_mocetinostat
GSM1744393_gefitinib	GSM1737642_mocetinostat	GSM1745149_GDC-0980
GSM1742645_gefitinib	GSM1738308_entinostat	GSM1737409_NVP-BGT226
GSM1746613_enzastaurin	GSM1742797_palbociclib*	GSM1737410_NVP-BGT226
GSM1742878_dasatinib*	GSM1739435_belinostat*	GSM1741767_vorinostat*

sion levels between 69 prostate cancer patients and 18 normal patients using Affymetrix Human Genome U133A 2.0 Array. This dataset is available in GEO (GSE6956) [277].

Gold standard: Nilutamide is the FDA-approved drug for the treatment of prostate cancer. This antiandrogen drug is included in our list of input drugs from Connectivity Map. Prostate cancer mostly depends on the androgen for the growth and survival. Nilutamide is known to block the action of androgens of adrenal and testicular origin that stimulate the growth of the normal and malignant prostatic tissue [128].

Tables 14–16 show the results obtained on 3 prostate cancer datasets using: i) the proposed approach, ii) the drug-disease approach, and iii) the anti-correlation approach.

We selected the top 5% drugs from the drugs ranked lists obtained by applying the proposed approach on 3 prostate cancer datasets. As shown in Table 13, these drugs are

Table 11: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE11969-squamous cell carcinoma dataset (the top 10 drugs). The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach with drug-disease and anti-correlation approaches are 0.001 and 0.0003, respectively. Drugs highlighted with green are FDA-approved for the treatment of NSCLC. The * denotes the drugs that are currently FDA-approved but for other indications.

GSE11969-squamous cell carcinoma		
Proposed	Drug-disease	Anti-correlation
GSM1741743_sirolimus*	GSM1739358_mocetinostat	GSM1737411_NVP-BGT226
GSM1741104_sunitinib*	GSM1738326_mocetinostat	GSM1740576_BI-2536
GSM1740080_sunitinib*	GSM1746780_erlotinib*	GSM1745149_GDC-0980
GSM1738326_mocetinostat	GSM1742797_palbociclib*	GSM1745194_NVP-BEZ235
GSM1746613_enzastaurin	GSM1737642_mocetinostat	GSM1739358_mocetinostat
GSM1739358_mocetinostat	GSM1742706_alvocidib	GSM1741767_vorinostat*
GSM1744393_gefitinib	GSM1745912_neratinib*	GSM1737410_NVP-BGT226
GSM1742878_dasatinib*	GSM1737967_entinostat	GSM1737409_NVP-BGT226
GSM1740081_sunitinib*	GSM1738308_entinostat	GSM1745191_NVP-BEZ235
GSM1740917_saracatinib	GSM1737624_entinostat	GSM1741769_sirolimus*

Table 12: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE32863 dataset (the top 10 drugs). The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach with drug-disease and anti-correlation approaches are 0.007 and 0.007, respectively. Drugs highlighted with green are FDA-approved for the treatment of NSCLC. The * denotes the drugs that are currently FDA-approved but for other indications.

GSE32863-adenocarcinoma		
Proposed	Drug-disease	Anti-correlation
GSM1741743_sirolimus*	GSM1738326_mocetinostat	GSM1738326_mocetinostat
GSM1738326_mocetinostat	GSM1739358_mocetinostat	GSM1739453_decitabine*
GSM1741104_sunitinib*	GSM1743209_nintedanib*	GSM1746780_erlotinib*
GSM1740080_sunitinib*	GSM1746780_erlotinib*	GSM1739358_mocetinostat
GSM1739358_mocetinostat	GSM1746881_mitoxantrone*	GSM1742878_dasatinib*
GSM1742878_dasatinib*	GSM1746893_radicicol	GSM1743210_nintedanib*
GSM1744393_gefitinib	GSM1742797_palbociclib*	GSM1742797_palbociclib*
GSM1740301_ponatinib*	GSM1740576_BI-2536	GSM1742706_alvocidib
GSM1740081_sunitinib*	GSM1740298_ponatinib*	GSM1744393_gefitinib
GSM1742552_linifanib	GSM1741767_vorinostat*	GSM1740576_BI-2536

ranked according to the scores computed by the systematic method from the highest to the lowest. The score assigned to Nilutamide is 0. Drugs with the same scores are sorted based on their average ranks in drugs ranked lists.

Table 13: The top 5% drugs obtained from the result of our repurposing approach. These drugs are ranked based on the scores generated by the systematic method. The * denotes the drugs that are currently FDA-approved but for other indications. The score for Nilutamide, the FDA-approved drug for prostate cancer, is 0. Drugs with the same scores are sorted based on their average ranks in drugs ranked lists.

Drug	Score
Podophyllotoxin *	0.33
Nilutamide	0
Acetylsalicylic acid *	-0.33
Papaverine *	-0.33
Mefloquine *	-0.33
Vorinostat *	-0.33
Sirolimus *	-0.33
Alprostadil *	-0.33
Glibenclamide *	-0.33
Oxyphenbutazone (discontinued/withdrawn)	-0.33
Phenelzine *	-0.33
Methylergometrine *	-0.33
Parthenolide	-0.33
Primaquine *	-0.33
Phenoxybenzamine *	-0.67
Etoposide *	-0.67
Captopril *	-0.67
Trichostatin A	-0.67
Fluorometholone *	-1

Proposed candidates: In this case study, we chose Podophyllotoxin (p = 0.2), Acetylsalicylic acid (p = 0.2), Papaverine (p = 0.01), Mefloquine (p = 0.03), Vorinostat (p = 0.1), and Sirolimus (p = 0.06) for further evaluations.

Podophyllotoxin is a natural product found in podophyllin resin from the roots of podophyllum plants. Podophyllotoxin and its derivatives, including Deoxypodophyllotoxin, are reported to have significant anti-tumor effects in a number of cancers [25, 48, 53, 88, 137, 159]. A recent study demonstrated that Deoxypodophyllotoxin inhibits the cell proliferation and induces the cell apoptosis in human prostate cancer cells through the Akt/p53/Bax/PTEN signaling pathway, suggesting that Deoxypodophyllotoxin could be used as a novel chemotherapeutic drug for human prostate cancer [109].

Acetylsalicylic acid (Aspirin) is a nonsteroidal anti-inflammatory drug that is used for the temporary relief of different forms of pain, and the inflammation associated with various conditions. It is also indicated to decrease the risk of death and myocardial infarction in patients with chronic coronary artery disease. Recent findings confirm that the long duration regular Aspirin use modestly reduces the risk of prostate cancer [52, 97, 118, 160, 188, 229, 249]. Aspirin is also reported to affect the proliferation, apoptosis, resistance and metastasis of prostate cancer cell lines, suggesting the further evaluation of the signaling cascades activated by Aspirin in order to improve diagnosis, prognosis and treatment of prostate cancer [29]. According to these findings, Aspirin can be used for both prevention and treatment purposes in prostate cancer (ClinicalTrials.gov IDs: NCT02757365, NCT03103152, NCT02804815).

Papaverine is a nonxanthine phosphodiesterase inhibitor that is indicated for the relief of the cerebral and peripheral ischemia. It induces morphologic differentiation and suppresses the proliferation of human prostate cancer cell [89]. Papaverine is reported to have antitumor effects in prostate cancer by inducing significant, highly selective and dose-dependent cytotoxic effects in cancer cells [112, 236].

Mefloquine (MQ) is a prophylactic anti-malarial drug which acts as a blood schizonticide and can be a potential treatment for prostate cancer. Recent findings indicate that MQ has anticancer effects in PC3, which is the most commonly used prostate cancer cell line [294, 295]. MQ has been reported to be potent in killing cancer cells in vitro, sug-

Table 14: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE26910 dataset (the top 10 drugs). The ranks of Nilutamide, the FDA-approved drug for prostate cancer, in the proposed approach, drug-disease and anti-correlation approaches results are 9, 13, and 63 , respectively. Drug highlighted with green is FDA-approved for the treatment of prostate cancer. The * denotes the drugs that are currently FDA-approved but for other indications.

GSE26910		
Proposed	Drug-disease	Anti-correlation
mefloquine_5724 *	luteolin_3041	vorinostat_6179 *
mefloquine_2210 *	etoposide_1626 *	parthenolide_5530
podophyllotoxin_2540	vorinostat_6939 *	parthenolide_2885
vorinostat_6179 *	phenoxybenzamine_5248 *	tanespimycin_2666
phenoxybenzamine_5613 *	puromycin_3310	phenoxybenzamine_5248 *
oxyphenbutazone_6844	ciclopirox_2456 *	phenoxybenzamine_5613 *
parthenolide_2885	vorinostat_6179 *	doxazosin_3024 *
parthenolide_5530	anisomycin_1304 *	mycophenolic acid_2857 *
nilutamide_5362	lycorine_3808	etoposide_3241 *

gested as the chemotherapeutic agent for treatment of glioblastoma and breast cancer cells [85, 247].

Vorinostat is a histone deacetylase (HDAC) inhibitor approved by FDA for the treatment of patients with cutaneous T-cell lymphoma (CTCL). Inhibition of the HDAC has resulted in decreasing the tumor growth and reducing cell proliferation in prostate cancer, suggesting that Vorinostat could be a potential drug for treatment of prostate cancer [32, 39, 66, 130] (ClinicalTrials.gov IDs: NCT00330161, NCT00589472).

The initial preclinical and clinical studies show that the mTOR inhibition **Sirolimus** can be useful in treating patients with prostate cancer [9, 40, 55, 114, 215]. Sirolimus and its combination with other drugs are undergoing clinical trials in treatment of patients with prostate caner (NCT00311623, NCT02565901).

Table 15: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE6919 dataset (the top 10 drugs). The ranks of Nilutamide, the FDA-approved drug for prostate cancer, in the proposed approach, drug-disease and anti-correlation approaches results are 7, 81, and 129, respectively. Drugs highlighted with green are FDA-approved for the treatment of prostate cancer. The * denotes the drugs that are currently FDA-approved but for other indications.

GSE6919		
Proposed	Drug-disease	Anti-correlation
papaverine_1755 *	alvespimycin_1638	doxorubicin_3291 *
vorinostat_6179 *	daunorubicin_4983 *	doxorubicin_5671 *
etoposide_3241 *	tanespimycin_986	daunorubicin_4983 *
alprostadil_2938 *	doxorubicin_5671 *	mitoxantrone_3232 *
podophyllotoxin_2540*	alvespimycin_993	rifabutin_3873 *
methylergometrine_1607	mitoxantrone_3232 *	alvespimycin_1638
nilutamide_5362	etoposide_3241 *	alvespimycin_993
fluorometholone_6247 *	parthenolide_5530	oxyphenbutazone_6844
colchicine_1598 *	ciclopirox_3317 *	mitoxantrone_5354 *
acetylsalicylic acid_1042 *	mitoxantrone_5354 *	vorinostat_6939 *

Table 16: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE6956 dataset (the top 10 drugs). The ranks of Nilutamide, the FDA-approved drug for prostate cancer in the proposed approach, drug-disease and anti-correlation approaches results are 15, 141, and 90, respectively. The * denotes the drugs that are currently FDA-approved but for other indications.

GSE6956		
Proposed	Drug-disease	Anti-correlation
acetylsalicylic acid_1042 *	phenelzine_4360 *	phenelzine_4360 *
captopril_1988 *	rifabutin_4349 *	rifabutin_4349 *
sirolimus_1080 *	trichostatin A_5017	primaquine_4845 *
glibenclamide_1546 *	captopril_1988 *	norfloxacin_7283 *
phenelzine_4360 *	ambroxol_6719 *	flunixin_2552
sirolimus_987 *	metaraminol_2298 *	captopril_1988 *
trichostatin A_5017	sirolimus_1080 *	sirolimus_987 *
primaquine_4845 *	picrotoxinin_2161 *	trichostatin A_5017
paclitaxel_6720	cyproheptadine_6740 *	calmidazolium_906
ajmaline_1749 *	primaquine_4845 *	ambroxol_6719 *

2.5.4 Breast cancer

We obtained six breast cancer datasets, GSE1299 [175] and GSE28645 [278], and GSE65194 (subtypes: Her2, luminalA, luminalB, triple negative) [165, 166, 170] from Gene Expression Omnibus (GEO). Datasets GSE1299 and GSE65194 (Her2, luminalA, luminalB, triple negative) consist of two groups of samples such as disease and control, while the dataset GSE28645 is a gene expression dataset that consists of two groups of samples: treated (by tamoxifen) and untreated. It is well-known that choices of the treatment and the ultimate success for breast cancer highly depend on its specific type [86], that is categorized as:

- Hormone receptor positive (estrogen and/or progesterone receptor positive) or hormone receptor negative (estrogen and/or progesterone receptor negative)
- Human epidermal growth factor receptor (HER2/neu) positive or HER2/neu negative
- Triple negative (all estrogen receptor, progesterone receptor, and HER2/neu are negative)

Other factors that affect the prognosis and treatment options include: stage of the cancer, levels of estrogen receptor, progesterone receptor, or HER2/neu in the tumor tissue, the growth rate of the tumor, the recurrence rate, patient's age, and menopausal status.

Gold standard: Fulvestrant, Paclitaxel, Methotrexate are FDA-approved drugs for the treatment of breast cancer. These drugs are included in our list of input drugs from Connectivity Map. Table 17 represents the target genes and activity of these drugs, obtained from KEGG and Drugbank [287]. Tables 19–24 show the results obtained on 6 breast cancer datasets using: i) the proposed approach, ii) the drug-disease approach, and iii) the anti-correlation approach.

We selected the top 5% drugs from the drugs ranked lists obtained by applying the proposed approach on 6 breast cancer datasets. These drugs are ranked according to the scores computed by the systematic method from the highest to the lowest. Table 18 shows the rank-ordered list of such drugs according to their score.

Proposed candidates: For this disease, we propose Captopril (p = 0.001), Glibenclimiade (p = 0.0009), Fluorometholone (p = 0.005), Etoposide (p = 0.01), Colchicine (p = 0.001), and Tretinoin (p = 0.0009) as repurposing candidates for treatment of breast cancer.

Captopril is indicated for treatment of hypertension, congestive heart failure, and kidney problems caused by diabetes. Recent clinical studies confirm the potential antineoplastic effect of Captopril in cancer [123, 129, 144, 225, 230]. The phase I/II clinical trial (ClinicalTrials.gov ID: NCT00086723) evaluates the activity of Captopril and the tissue plasminogen activator (a blood factor/protein orchestrating the breakdown of blood clot) in treating patients with progressive metastatic cancer. Specifically, Captopril is proven to play a role in prevention and regression of the tamoxifen-induced resistance of breast cancer cell line MCF-7 [185], suggesting that it can be used in combination with Tamoxifen to overcome such resistance.

Glibenclimiade is an antidiabetic drug that is used as an adjunct to diet and exercise for treatment of patients with type 2 diabete. Glibenclamide is proven to be a tumor growth inhibitor [2, 202, 213, 224, 299]. It is considered as a promising antitumor drug in several cancers, including breast cancer. In particular, the cytostatic effect of Glibenclimiade by inducing G0/G1 arrest has been clearly demonstrated in MDA-MB-231 cells. Additionally, the study of its effect in combination with Doxorubicin suggests the novel role of Glibenclimiade as an adjuvant in breast cancer treatment [190].

Fluorometholone and **Clobetasol** are in the family of glucocorticoids (GCs). GCs have shown some modest benefits in treatment of breast cancer. However, their underlying mechanism in breast cancer is not well-understood [127, 132, 155]. GCs are also used as an adjuvant during chemotherapy or radiotherapy to reduce the side effects in cancer treatment [290].

Etoposide is approved by FDA for the treatment of refractory testicular tumors, and usually used in combination with other chemotherapeutic agents. It is also used as the first line treatment in small cell lung cancer patients. The positive therapeutic effect of Etoposide in patients with breast cancer is experimentally validated by clinical studies [18, 306] (ClinicalTrials.gov identifiers: NCT01492556, NCT00026949, and NCT01589159).

The histone deacetylase (HDAC) inhibitor **Trichostatin A (TSA)** is another drug we suggest for treatment of breast cancer. It is used as an antifungal antibiotic that is found to be useful both as the single agent and in combination with other agents in cancer treatment [119, 120, 134, 147, 182, 223, 257]. Current studies confirm the potent antitumor activity of TSA against breast cancer [7, 221, 272, 293, 298].

Colchicine is found in crocuses and primarily indicated to treat gout. It has been also used for treatment of familial mediterranean fever. A 12-year study in male patients with gout shows that patients who used Colchicine had a significantly lower risk of cancers than patients who never used Colchicine [145]. Another study in mice models shows that Colchicine can induce immunogenic cell death in tumor cells, suggesting the future clinical evaluation for Colchicine as a cancer vaccine [284]. Colchicine is reported to have an anticancer effect on human gastric cancer cell lines [156]. In particular, a recent study indicates that it can inhibit proliferation of the breast cancer MCF-7 cells and induce cell apoptosis, where the intensity of the effect depends on the time and dosage [255].

Tretinoin, all-trans-retinoic acid (ATRA), is the FDA-approved drug for the treatment of acne, photodamaged skin, and keratinization disorders. It is also used to treat acute promyelocytic leukemia (APL). The usefulness of ATRA in treatment of breast cancer has been independently validated in study by Bhat-Nakshatri et al. [27]. Moreover, anti-proliferative, cyto-differentiating and apoptotic effects of ATRA are demonstrated in [33, 82, 256], suggesting the effectiveness of ATRA in treatment of breast cancer tumors with high retinoic acid receptor alpha (RARα) / retinoic acid receptor gamma (RARγ) ratios. Estrogen receptor-positive and Her2/neu-positive breast cancers are two subtypes of breast cancer that can be optimal targets for ATRA [157, 199].

Table 17 represents the target genes and activity of these drugs, obtained from KEGG and Drugbank [287].

Table 17: The FDA-approved drugs for breast cancer.

Drug	Target genes	Activity
Fulvestrant	*ESR (1,2)*	Estrogen receptor antagonist
Methotrexate	*DHFR*	Antimetabolite
Paclitaxel	*BCL2, TUBB1, NR1I2, MAP (2,4), MAPT*	Tubulin depolymerization inhibitor

Tables 19–24 show the results obtained on 6 breast cancer datasets using: i) the proposed approach, ii) the drug-disease approach, and iii) the anti-correlation approach. The FDA-approved drugs for breast cancer are highlighted in green. Interestingly, all two in-

stances of Fulvestrant that are all exposures of the same cell line (MCF7) (with the same dosage) are highly ranked by our approach in all datasets. MCF7 cell line is an ideal model for hormone therapy that was established in 1973 at the Michigan Cancer Foundation [253].

We selected the top 5% drugs from the drugs ranked lists obtained by applying the proposed approach on 6 breast cancer datasets. These drugs are ranked according to the scores computed by the systematic method from the highest to the lowest. Table 18 shows the rank-ordered list of such drugs according to their score.

Table 18: The top 5% drugs obtained from the result of our repurposing approach. These drugs are ranked based on the scores generated by the systematic method. The * denotes the drugs that are currently FDA-approved but for other indications. Fulvestrant, Paclitaxel, and Methotrexate highlighted with green, are FDA-approved for the treatment of breast cancer. The sum of the scores assigned to these drugs is 0. Drugs with the same scores are sorted based on their average ranks in drugs ranked lists. In this analysis, the scores computed by our systematic method can be further normalized based on dividing each score by the total number of FDA-approved drug instances for the disease of interest. Therefore, the scores calculated for each drug across different diseases will be in the same range (-1,1).

Drug	Score
Fulvestrant	1.33
Paclitaxel	0
Captopril *	-0.33
Methotrexate	-1.33
Glibenclamide *	-2
Fluorometholone *	-2.33
Clobetasol	-2.67
Trichostatin A	-2.83
Etoposide *	-3.33
Colchicine *	-3.67
Tretinoin *	-3.67
Alvespimycin	-3.67
Resveratrol	-3.67
Methylergometrine	-4

Table 19: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE65194-Her2 dataset (the top 10 drugs). The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach with drug-disease and anti-correlation approaches are 0.01 and 0.02, respectively. Drugs highlighted with green are FDA-approved for the treatment of breast cancer. The * denotes the drugs that are currently FDA-approved but for other indications. Such drugs can be used off-label.

GSE65194-Her2		
Proposed	Drug-disease	Anti-correlation
captopril_1988 *	glibenclamide_1546	glibenclamide_1546
paclitaxel_6720	danazol_2038 *	cimetidine_1884 *
fulvestrant_1630	cimetidine_1884 *	danazol_2038 *
fulvestrant_985	domperidone_2655	ajmaline_1749 *
etoposide_3241 *	trichostatin A_5017	domperidone_2655
captopril_4585 *	ipratropium bromide_1769	ipratropium bromide_1769
ipratropium bromide_1769	etoposide_3241 *	acepromazine_1777
metoclopramide_2353 *	ajmaline_1749 *	nilutamide_5362
domperidone_2655	methotrexate_5000	genistein_5232
methotrexate_5000	resveratrol_841 *	captopril_1988 *

Table 20: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE65194-LuminalA dataset (the top 10 drugs).The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach with drug-disease and anti-correlation approaches are 0.03 and 0.04, respectively. Drugs highlighted with green are FDA-approved for the treatment of breast cancer. The * denotes the drugs that are currently FDA-approved but for other indications. The proposed approach was the only one who was able to rank the FDA-approved drugs in the top 10.

GSE65194-LuminalA		
Proposed	Drug-disease	Anti-correlation
fulvestrant_985	glibenclamide_1546 *	domperidone_2655
fulvestrant_1630	domperidone_2655	glibenclamide_1546 *
captopril_1988 *	cimetidine_1884 *	cimetidine_1884 *
fluorometholone_6247 *	danazol_2038 *	ipratropium bromide_1769
glibenclamide_1546 *	ipratropium bromide_1769	danazol_2038 *
captopril_4585 *	trichostatin A_5017	nilutamide_5362
paclitaxel_6720	nilutamide_5362 *	ajmaline_1749 *
vorinostat_6939 *	ethosuximide_1433 *	genistein_5232
cimetidine_1884 *	ajmaline_1749 *	acepromazine_1777
trichostatin A_5017	genistein_5232	ethosuximide_1433 *

Table 21: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE65194-LuminalB dataset (the top 10 drugs). The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach with drug-disease and anti-correlation approaches are 0.02 and 0.01, respectively. Drugs highlighted with green are FDA-approved for the treatment of breast cancer. The * denotes the drugs that are currently FDA-approved but for other indications.

GSE65194-LuminalB		
Proposed	Drug-disease	Anti-correlation
fulvestrant_985	glibenclamide_1546 *	glibenclamide_1546 *
fulvestrant_1630	cimetidine_1884 *	cimetidine_1884 *
paclitaxel_6720	danazol_2038 *	danazol_2038 *
captopril_1988 *	ajmaline_1749 *	ajmaline_1749 *
trichostatin A_5017	trichostatin A_5017	domperidone_2655
phenelzine_4360 *	domperidone_2655	ipratropium bromide_1769
fluorometholone_6247 *	etoposide_3241 *	fluorometholone_6247 *
valproic acid_2700 *	ipratropium bromide_1769	sirolimus_1080 *
etoposide_3241 *	sirolimus_1080 *	acepromazine_1777
glibenclamide_1546 *	methotrexate_5000	nilutamide_5362

Table 22: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE65194-Triple Negative dataset (the top 10 drugs). The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach with drug-disease and anti-correlation approaches are 0.03 and 0.007, respectively. Drugs highlighted with green are FDA-approved for the treatment of breast cancer. The * denotes the drugs that are currently FDA-approved but for other indications.

GSE65194-Triple Negative		
Proposed	Drug-disease	Anti-correlation
captopril_1988 *	glibenclamide_1546 *	glibenclamide_1546 *
paclitaxel_6720	danazol_2038 *	danazol_2038 *
etoposide_3241 *	cimetidine_1884 *	ajmaline_1749 *
clobetasol_6835	ajmaline_1749 *	cimetidine_1884 *
methotrexate_5000	methotrexate_5000	ipratropium bromide_1769
fulvestrant_985	resveratrol_841 *	domperidone_2655
glibenclamide_1546 *	ipratropium bromide_1769	acepromazine_1777
fulvestrant_1630	trichostatin A_5017	etoposide_3241 *
captopril_4585 *	acepromazine_1777	nilutamide_5362
domperidone_2655	wortmannin_1023 *	methotrexate_5000

Table 23: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE28645 dataset (the top 10 drugs). The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach with drug-disease and anti-correlation approaches are 0.05 and 0.07, respectively. Drugs highlighted with green are FDA-approved for the treatment of breast cancer. The * denotes the drugs that are currently FDA-approved but for other indications.

GSE28645		
Proposed	Drug-disease	Anti-correlation
fulvestrant_985	paclitaxel_6720	phenelzine_4360 *
fulvestrant_1630	captopril_1988 *	valproic acid_1181 *
methylergometrine_1607	phenelzine_4360 *	paclitaxel_6720
alvespimycin_993	fluphenazine_6954 *	rosiglitazone_4457 *
colchicine_1598 *	clomipramine_4487 *	captopril_1988 *
captopril_1988 *	rosiglitazone_4457 *	troglitazone_4456 *
vorinostat_6939 *	genistein_5232	quercetin_2499
captopril_4585 *	astemizole_6807 *	hyoscyamine_1424 *
sirolimus_1080 *	chlorpromazine_5074 *	clomipramine_4487 *
trichostatin A_5017	cimetidine_1884 *	genistein_1176

Table 24: A comparison between the results of three approaches: proposed, drug-disease, anti-correlation using GSE1299 dataset (the top 10 drugs). The *p*-values for Wilcoxon rank sum test comparing the results of the proposed approach with drug-disease and anti-correlation approaches are 0.007 and 0.02, respectively. Drugs highlighted with green are FDA-approved for the treatment of breast cancer. The * denotes the drugs that are currently FDA-approved but for other indications.

GSE1299		
Proposed	Drug-disease	Anti-correlation
methotrexate_5419	etoposide_3241 *	etoposide_3241 *
resveratrol_841 *	ciclopirox_3317 *	valproic acid_1181 *
methotrexate_5000	resveratrol_841 *	rifabutin_4349 *
fulvestrant_985	valproic acid_1181 *	methotrexate_5419
tretinoin_6170 *	rifabutin_4349 *	resveratrol_841 *
phenelzine_4360 *	ivermectin_2213 *	rifabutin_3873 *
fulvestrant_1630	oxyphenbutazone_6844	ciclopirox_3317 *
tretinoin_1548 *	methotrexate_5419	prochlorperazine_5212 *
troglitazone_4456 *	rifabutin_3873 *	vorinostat_6939 *
rosiglitazone_4457 *	wortmannin_1023 *	phenelzine_4360 *

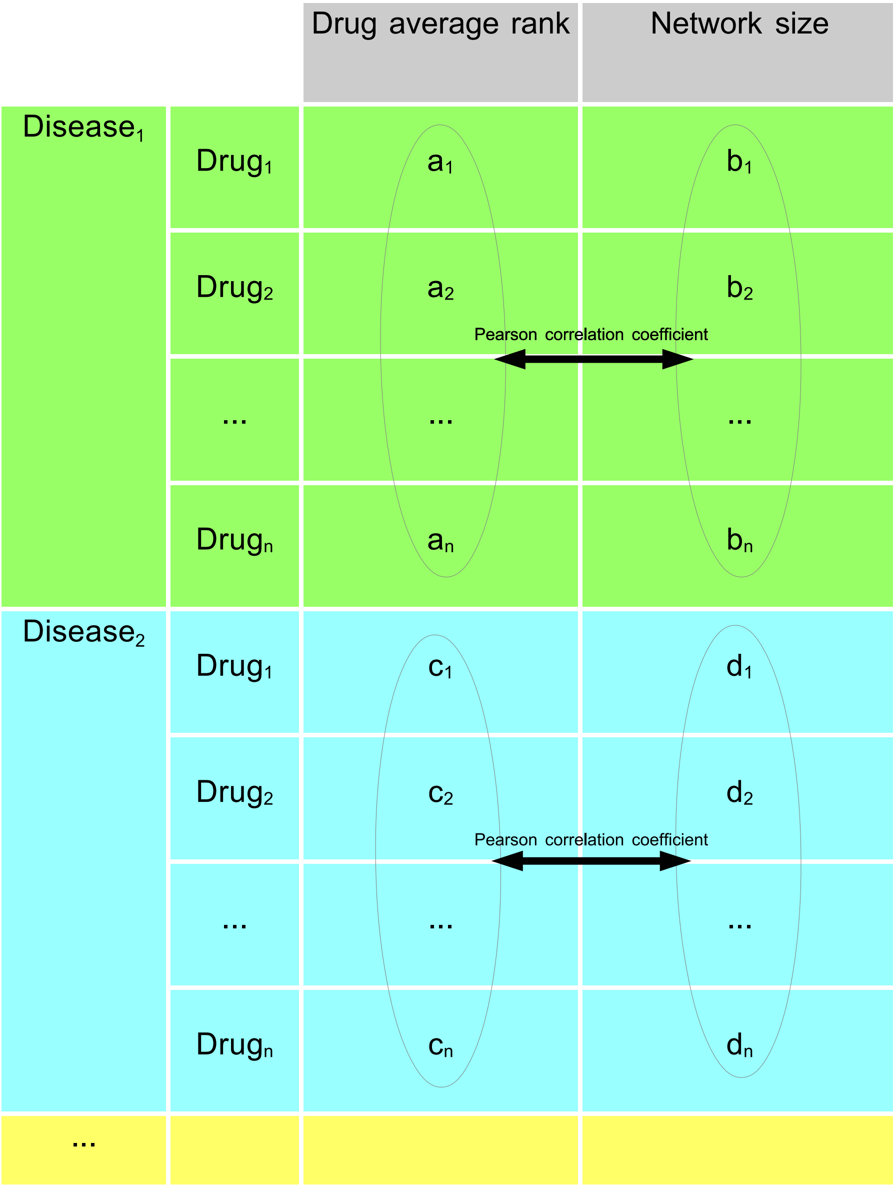
2.5.5 The relationship between pathway size and prediction accuracy

The prediction accuracy of the proposed approach does not depend on size of the KEGG pathways, nor on the size of the drug-disease network (DDN) constructed for each drug. In other words, a drug with a larger drug-disease network is not more likely to be ranked higher just because of the size of its network.

In order to prove this, we calculated the Pearson correlation coefficient between the drug's average ranks (obtained from applying our approach on disease datasets) and their networks size (the number of nodes) for each disease (Figure 10 and 11). If the ranking of a drug depended on the size of its drug-disease network (e.g. the top ranked drug would have the largest network, the second drug would have the second largest network, etc.), then there would be a very strong correlation between the ranks and the network sizes. In fact, the computed Pearson correlation coefficients between the ranks and the network sizes were: 0.004 for IPF, 0.25 for NSCLC, -0.10 for breast cancer, and -0.14 for prostate cancer. Both the low absolute values of these correlation coefficients as well as the fact that two of them are positive and while the other two are negative show no indication of any dependency between the pathway sizes and the predictions.

2.6 Summary

We presented a systems biology approach to discover new uses of existing FDA-approved drugs. We take advantage of known knowledge of disease-related genes, drug targets information, and signaling pathways to discover drugs with the potential desired effects on the given disease. We estimated a network of genes potentially perturbed by drugs and integrate this network with drug and disease gene expression signatures to conduct a more



		Drug average rank	Network size
Disease$_1$	Drug$_1$	a_1	b_1
	Drug$_2$	a_2	b_2
	...	...	...
	Drug$_n$	a_n	b_n
Disease$_2$	Drug$_1$	c_1	d_1
	Drug$_2$	c_2	d_2
	...	...	...
	Drug$_n$	c_n	d_n
...			

Pearson correlation coefficient
Pearson correlation coefficient

Figure 10: The relationship between pathway size and prediction accuracy. In order to investigate the relationship between pathway size and prediction accuracy, we calculated the Pearson correlation coefficient between the drug's average ranks and their networks size for each disease.

powerful analysis at system level. To evaluate the proposed approach for drug repurposing, four different diseases (IPF, NSCLC, prostate cancer, and breast cancer) were analyzed using 3 approaches: proposed, drug-disease, and anti-correlation. For each disease, there

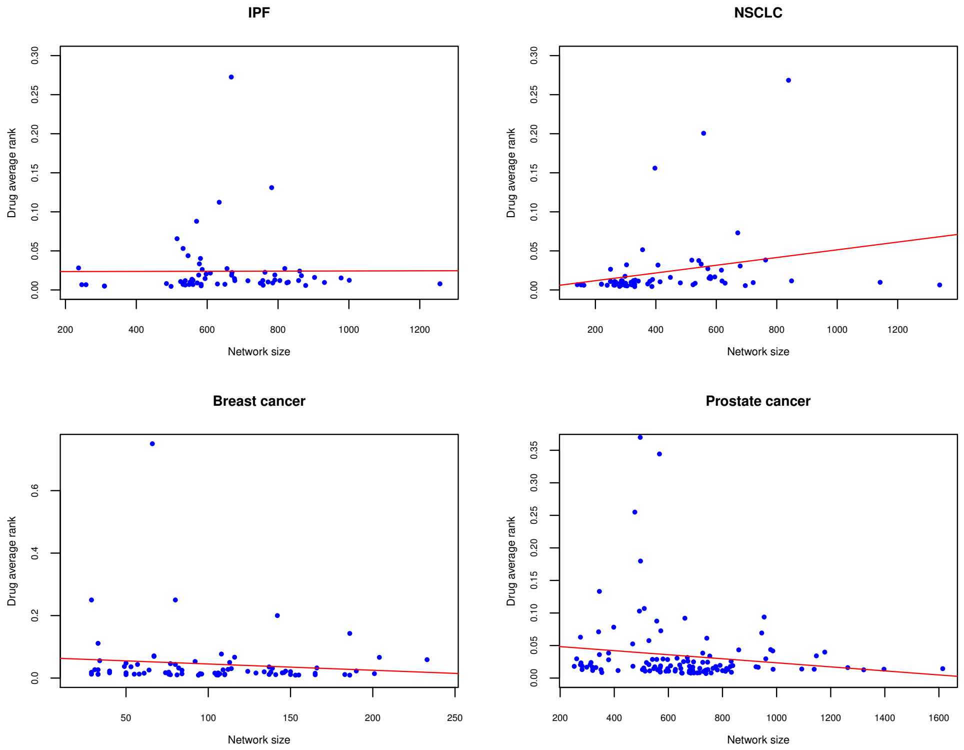


Figure 11: The scatter plots of the drug relative average rank vs. network size for four diseases: IPF, NSCLC, breast cancer, and prostate cancer. The scatter plots of the drug relative average rank vs. network size for four diseases: IPF, NSCLC, breast cancer, and prostate cancer. The Pearson correlation coefficients between the ranks and the network sizes were: 0.004 for IPF, 0.25 for NSCLC, -0.10 for breast cancer, and -0.14 for prostate cancer. The low correlations show there is no indication of a dependency between the pathway size and the predicted ranks.

is at least one FDA-approved drug that is used to treat that disease in our input drug data. The already FDA-approved drugs for a given disease are considered as the gold standard because such drugs successfully passed all the preclinical and clinical trials for that disease and were demonstrated to be efficacious in each disease. The approach was validated by its ability to identify drugs that are already approved by FDA for these conditions. We provided evidences that support the usefulness of the proposed candidates in treatment of IPF, NSCLC, prostate cancer, and breast cancer. In all diseases, the proposed approach was able to rank highly the approved drugs for the given conditions. This is in contrast with the Sirota et al. (drug-disease) and anti-correlation approaches which were not able to retrieve any of the FDA approved drugs at the top of their respective rankings. For many of the proposed repurposing candidates, there is significant preliminary evidence, as well as a number of clinical trials in progress.

CHAPTER 3 CELL TYPE IDENTIFICATION

3.1 Problem statement

Recent advances in single-cell RNA-Seq (scRNASeq) techniques have provided transcriptomes of the large numbers of individual cells (single-cell gene expression data) [61, 79, 99, 126, 168, 186, 194, 266, 280]. In particular, analyzing the diversity and evolution of single cancer cells can enable the advances in early cancer diagnosis, and ultimately choosing the best strategy for cancer treatment [149, 231, 244]. Furthermore, one important analysis on scRNASeq is the identification of cell types that can be achieved by performing an unsupervised clustering method on transcriptome data [12, 13, 75, 177, 265, 291, 304, 308].

3.2 Overview of existing approaches

Clustering algorithms such as k-means and density-based spatial clustering of applications with noise (DBSCAN) [78] can identify groups of cells given the single-cell gene expression data. However, clusters obtained by these algorithms might not be robust. Such algorithms require non-intuitive parameters [12]. For instance, given the number of clusters, k-means iteratively assigns data points (cells) to the nearest centroids (cluster center), and recomputes the centroids based on the predefined number of clusters. This algorithm starts with the randomly chosen centroids. Thus, the result of the algorithm depends on the number of clusters (in DBSCAN, the maximum distance between the two data points in the same neighborhood should be determined) and the number of runs.

Another challenge comes from the high dimensionality of data, known as “curse of dimensionality”. Identifying the accurate clusters of data points based on the measured distances between the pairs of data points may fail since those data points become more similar when they are represented in a higher dimensional space [12, 22]. One approach to deal with the curse of high dimensionality is projecting data into a lower dimensional space, known as dimensionally reduction. In this approach, the data is represented in a lower dimensional space while the characteristic(s) (e.g similarities between the data points) of the original data is preserved. Several methods have used different techniques based on this concept (e.g. principal component analysis) to determine the cell types [19, 37, 162, 208, 240]. Another approach to deal with this challenge is feature selection, i.e. eliminating some of the features (genes) that are not informative [96]. In the following, we provide a brief overview of the related methods that identify the cell types based on the combination of approaches described above.

Methods SC3 [139] and Seurat [162] use a combination of feature selection, dimensionality reduction, and clustering algorithms to identify the cell types. Authors of SC3 use a consensus clustering framework that combines clustering solutions obtained by the spectral transformations and k-means clustering based on the complete-linkage hierarchical clustering. They first apply a gene filtering approach on the single-cell gene expression data to remove rare and ubiquitous genes/transcripts. Next, they compute the distance matrices (distance between the cells) using the Euclidean, Pearson, and Spearman metrics. They transform the distance matrices using either principal component analysis (PCA) [121], or by computing the eigenvectors of the associated graph Laplacian. Next, they perform a k-means clustering on the first d eigenvectors of the transformed distance matrices. Using the different k-means clustering results, they construct a consensus matrix that represents how often each pair of cells is clustered together. This consensus matrix is used as an input to a hierarchical clustering using a complete linkage and agglomeration strategies [70]. The clusters are inferred at the k-th level of hierarchy, where k is computed based on the Random Matrix Theory [201, 264]. The accuracy of SC3 is sensitive to the number of eigenvectors (d), chosen for the spectral transformation. The authors report that SC3 performs well when d is between 4% and 7% of the number of cells. The main advantage of SC3 is its high accuracy in identification of cell types. However, it is not scalable [138].

Seurat [162] is a graph-based clustering method that projects the single cell expression data into the two-dimensional space using the t-distributed stochastic neighbor embedding (t-SNE) technique [161]. Then, it performs the DBSCAN method [78] on the dimensionality-reduced single cell data. Seurat may fail to find the cell types in small datasets (low cell numbers) [139]. It is reported that this may be due to possible difficulties in estimating the densities when the number of data points is low.

RaceID [93] determines the cell types by performing a k-means clustering algorithm. In this method, the gap statistics is used to choose the number of clusters. RaceID does not perform well when the data does not contain rare cell populations but it appears to be the preferred methods when the aim is identification of rare types [12, 138, 151, 154].

SNN-Cliq [291] uses the shared nearest neighbor (SNN) concept, which considers the effect of the surrounding neighbor data points, to handle the high-dimensional data. The authors of SNN-Cliq compute the similarity between the pairs of data points (the similarity matrix) based on the Euclidean distance, referred as the primary similarity measure. Using the similarity matrix, they list the k-nearest neighbors (KNN) to each data point. They propose a secondary similarity measure that computes the similarity between two data points based on their shared neighborhoods. Consequently, an SNN graph is constructed based on the connectivity between the data points. Then, a graph-based clustering method is applied on the SNN graph in which nodes and weighted edges represent the data points and similarities between the data points, respectively. The main disadvantage of the graph-based methods such as SNN-Cliq is that scRNASeq data is not inherently graph-structured [12]. Therefore, the accuracy of these methods depends on the graph representation of scRNASeq data.

SINCERA [95] performs a hierarchical clustering on the similarity matrix that is computed using the centered Pearson's correlation. The average linkage approach is used as the default choice for the linkage. Consensus clustering [181, 286], tight clustering [268] and ward linkage [283] are provided as alternative clustering approaches. Users can choose a distance threshold or the number of clusters during the visual inspection when the hierarchical clustering is used for the cell cluster identification. SINCERA tends to identify many clusters which likely represent the same cell type [12].

One way to identify robust clusters of cells is to resample the cells/genes and compare the original clusters with the ones that are obtained by resampling [124]. In this thesis, in order to explore the strength of a pattern (cluster of cells) in the data, we analyze the sensitivity of that pattern against small changes in the data. The data is resampled by replacing a certain number of data points with the noise points from a noise distribution. Our hypothesis is that if there is a strong pattern in data, it will remain despite small perturbations [73]. Here, we develop a stable subtyping (clustering) method that employs the t-distributed stochastic neighbor embedding (t-SNE) [161] and k-means clustering to identify the cell types. We add noise and apply a bootstrap method [100, 101] to identify the stable clusters of cells. We use the Adjusted Rand Index (ARI) [113], adjusted mutual information (AMI) [274, 275], and V-measure [226] to evaluate the performance of the clustering result for datasets in which the true cell types are known. We compare the results of our method [206] with five other methods: RaceID [93], SC3 [139], SEURAT [162], SINCERA [95], and SNN-Cliq [291] using 8 real datasets with known cell types and 5 simulated datasets. The results of the different methods show that the proposed method performs better than the five methods across different datasets.

3.3 Data source

The goal of the proposed method is to identify the cell types present in a mixture of single cells. The input of the method is the single cell gene expression matrix ($\mathrm{M}_{gene\times cell}$) in which rows represent the genes and columns represent the cells. In the following we provide more detail about the input data and different steps of the proposed framework. The overall approach is shown in Figure 12.

The eight publicly available scRNA-seq datasets as well as the five simulation datasets were used in our analysis.

The list of eight single cell datasets we used in our analysis is summarized in Table 25. We obtained the processed data from Hemberg lab's website (https://hemberg-lab.github.io/scRNA.seq.datasets). Hemberg et al. [140] use the SingleCellExperiment Bioconductor S4 class [158] to store the data, and the scater package [172] for the quality control and plotting purposes. The normalized data is deposited as a SingleCellExperiment object (.RData file) and the cell type information is accessed in the cell_type1 column of the "colData" slot of this object. The gene expression values of the cells are organized as a matrix in which rows are cells and columns are the genes. In our analysis, genes (features) that are not expressed in any cells are removed. We did not filter any cell in this analysis.

We simulated four scRNA-seq datasets with varying degree of cluster separability using the splatter R package [307].The dataset $sim3$, consists of 3 subpopulations (1000 cells) with relative abundances 0.35, 0.30, and 0.35. The dataset $sim4$ includes 4 subpopulations (3000 cells) with relative abundances 0.15, 0.3, 0.2, and 0.35. The dataset $sim6$ includes 6 subpopulations (1000 cells) with relative abundances 0.3, 0.1, 0.1, 0.2, 0.2, and 0.1. Finally, the dataset $sim8$ consists of 8 subpopulations (2000 cells) with relative abundances 0.05, 0.1,0.1, 0.2, 0.2, 0.1, 0.15, and 0.1. We also used SPARSim R package [21] to generate a simulation dataset with 8 subpopulations (564 cells) from the real dataset Tung [269]. Among the eight real datasets, all but three (Klein [142], Patel [200], Treutlein [267]) are considered as 'gold standard' since the labels of the cells are known in a definitive way. Patel [200] and Treutlein [267] are referred as 'silver standard' by Kiselev et al. [139] since their cell labels are determined based on the computational methods and the authors' knowledge of the underlying biology.

Table 25: Single cell datasets. All the datasets, except Klein, Patel, and Treutlein are considered as “gold standard”. In the “gold standard” datasets, the cell types are clearly known. Klein, Patel, and Treutlein are referred as “silver standard' by [139] since the cell types are determined based on the computational methods and the authors' knowledge of the underlying biology.

Dataset	# cell type	Organism	#cell	Source	Reference
Biase	3	Mouse	49	Embryo development	[28]
Deng	10	Mouse	268	Embryo development	[65]
Goolam	5	Mouse	124	Embryo development	[87]
Klein	4	Mouse	2717	Embryo Stem Cells	[142]
Patel	5	Human	430	Tissues	[200]
Pollen	11	Human	301	Tissues	[210]
Treutlein	5	Mouse	80	Tissues	[267]
Yan	8	Human	124	Embryo development	[296]

3.4 Proposed framework

One way to identify robust clusters of cells is to resample the cells/genes and compare the original clusters with the ones that are obtained by resampling [124]. In the current thesis, in order to explore the strength of a pattern (cluster of cells) in the data, we analyze the sensitivity of that pattern against small changes in the data. The data is resampled by replacing a certain number of data points with the noise points from a noise distribution. Our hypothesis is that if there is a strong pattern in data, it will remain despite small perturbations [73]. Here, we develop a stable subtyping (clustering) method that employs the t-distributed stochastic neighbor embedding (tSNE) [161] and k-means clustering to identify the cell types. We add noise and apply a bootstrap method [100, 101] to identify the stable clusters of cells. We use the Adjusted Rand Index (ARI) [113] to evaluate the performance of the clustering result for datasets in which the true cell types are known. We compare the results of our method with five other methods: RaceID [93], SNN-Cliq [291], SINCERA [95], SEURAT [162], and SC3 [139] using eight datasets with known cell types. The ARIs computed on the results of the different methods show that the proposed method performs better than the five methods across different datasets.

3.4.1 Gene filtering

As shown in Figure 12A, we remove the genes/transcripts that are not expressed in any cell (expression value is zero in all cells). Such genes cannot provide useful information that can differentiate between cell types [11]. The result of performing the filtering method on the single cell gene expression matrix ($M_{gene \times cell}$) is used as the input to the second module of the proposed framework.

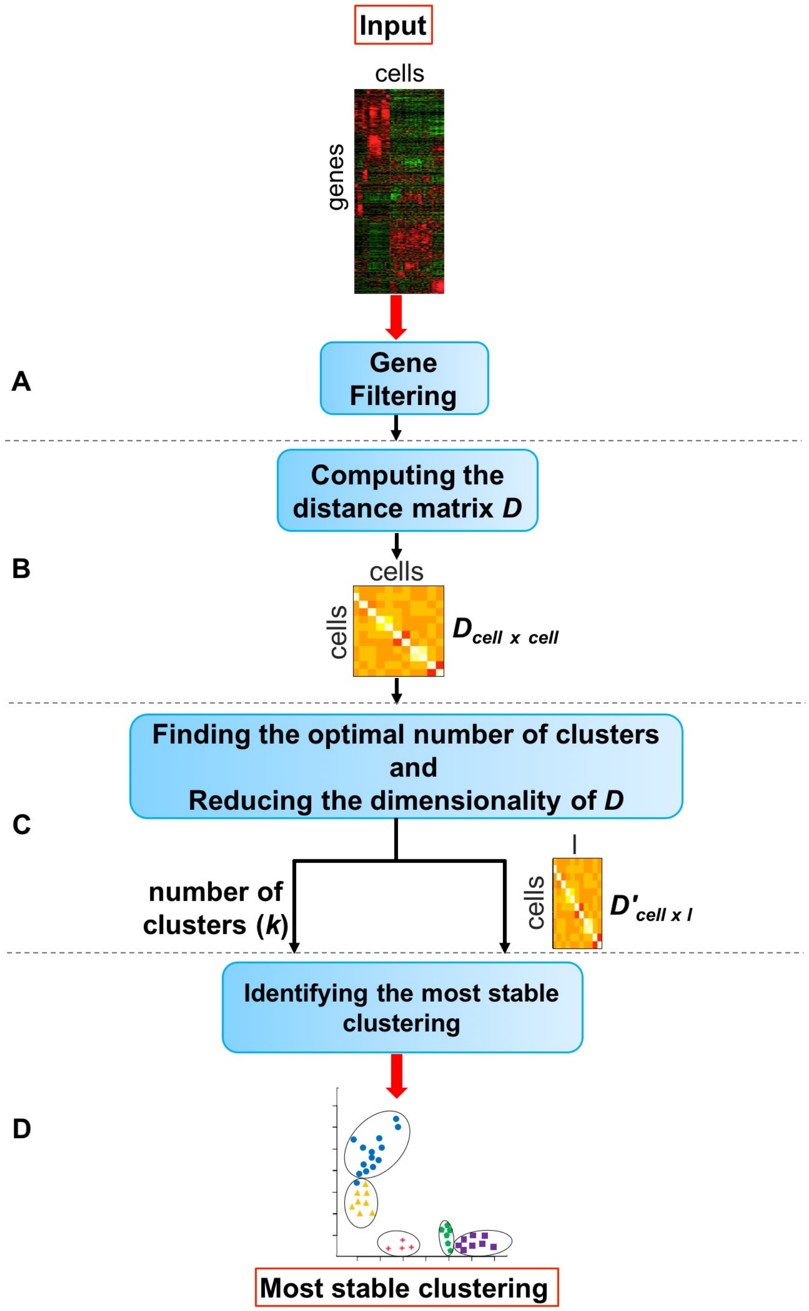

Input
cells
genes
A
Gene Filtering
Computing the distance matrix D
B
cells
cells
D cell x cell
Finding the optimal number of clusters and Reducing the dimensionality of D
C
number of clusters (k)
l
cells
D' cell x l
Identifying the most stable clustering
D
Most stable clustering

Figure 12: The overall workflow of the proposed method for the single cell identification [206]. Given the single cell gene expression matrix, **module (A)** eliminates the genes that are not expressed in any cell. Using the resulting matrix, **module (B)** computes the Euclidean distance between the cells. The output of this module is a distance matrix in which the rows and columns are the cells ($D_{cell \times cell}$). **Module (C)** reduces the dimensionality of the distance matrix using the t-distributed stochastic neighbor embedding (t-SNE) technique. In this module, an average silhouette method is employed to choose the optimal number of clusters k. Finally in **module (D)**, the lower-dimension distance matrix and the optimal number of clusters k obtained from **module (C)** are used as the input data to identify the most stable clustering of cells. Figure 13 shows the details of **module D**.

3.4.2 Measuring the dissimilarity between the cells

The distance between the cells is calculated using the Euclidean metric (Figure 12B). The output of this step is the distance (dissimilarity) matrix $D_{cell \times cell}$. We reduce the dimension of D by performing the t-distributed stochastic neighbor embedding (t-SNE) [10, 161], the nonlinear dimensionality reduction/visualization technique (Figure 12C). We will refer to the output as $D'_{cell \times l}$, where $2 \leq l \leq cell$. In this study, the number of dimensions is 2.

3.4.3 Identification of the optimal number of clusters

This section describes the third module of the proposed method (Figure 12C). In this analysis, the t-SNE is repeatedly (n=50) applied on the distance matrix $D_{cell \times cell}$ to obtain the dimensionality-reduced distance matrix $D'_{cell \times l}$. Each time, the optimal number of clusters is calculated based on the average silhouette method using the dimensionality reduced distance matrix D'. In order to find the optimal number of clusters k, the k-means clustering is applied on the D' matrix using a range value (default= 2:20), and the k that maximizes the average silhouette measure is selected. Finally, the average of the selected numbers k across different repeats $(n = 50)$ (rounded to the nearest integer) is considered as the final optimal number of clusters.

The silhouette evaluates the quality of that clustering based on how well its data points are clustered. A silhouette measure is assigned to each data point representing how close a data point is to its own cluster in comparison to other clusters. For each data point i, this measure is calculated as follows:

$$\text{s(i)} = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

where $a(i)$ is the average distance between the data point i and all other data points within the same cluster. $b(i)$ is the smallest average distance of i to all points in any other cluster of which i is not a member. $s(i)$ takes values from -1 to 1, where a high positive score shows that the given data point is well clustered (close to other points in its own cluster and far from points in the other clusters). Conversely, a high negative score shows that data point is poorly clustered.

3.4.4 K-means clustering based on the resampling method

This section describes the detail of the last module of the proposed method. As shown in Figure 13, using the dimensionality reduced distance matrix D' and the chosen number of clusters k from the previous step, we identify the most stable clustering by generating different clustering solutions ($clustering_i$ ($i \in [1..n]$)) and measure the stability of each clustering solution based on a resampling method. The stability measure assigned to each particular clustering (denoted as $clustering_i$) represents how often the k clusters belonging to that clustering are preserved when the input data (D') is resampled several times. The resampled datasets are generated from D' by randomly replacing 5% of data points (cells) with noise. These noisy datasets are then used as the input to k-means algorithm. Hence, several clusterings ($clustering_{i,j}$, $j \in [1..m]$) are generated from the resampled data (resampled versions of $clustering_i$).

In order to assess the stability of each cluster c in the $clustering_i$ (original clustering),

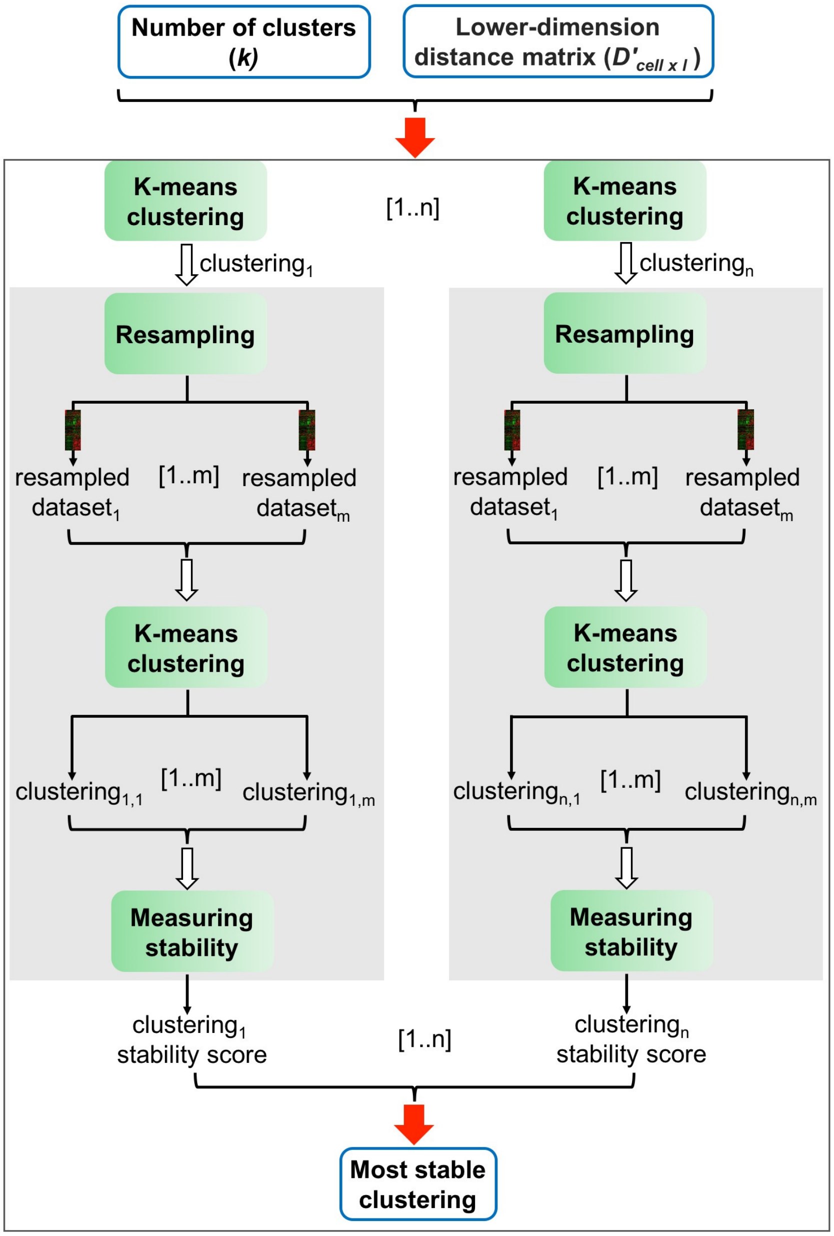


Figure 13: Identifying the most stable clustering. In this analysis, given the lower-dimension distance matrix $D'_{cell \times l}$ and the optimal number of clusters k, we calculate n different clusterings ($clustering_1, ..., clustering_n$) using the k-means clustering algorithm. Then, the stability of each clustering is assessed based on a resampling approach (grey box). A stability score is assigned to each clustering based on how often its clusters are recovered when the input data is perturbed (resampled). A clustering with the maximum stability score is selected as the final solution.

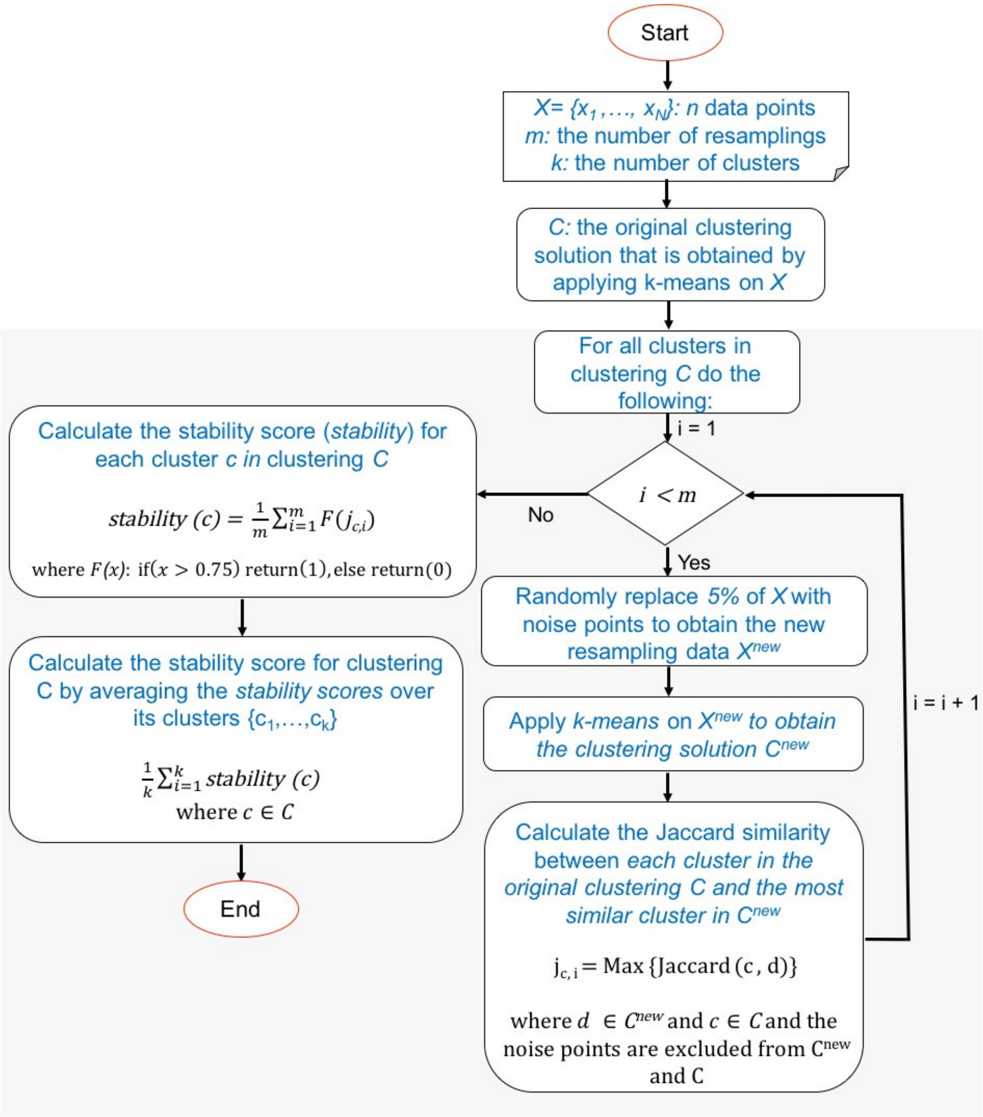

Start
$X = \{x_1, \ldots, x_N\}$: n data points
m: the number of resamplings
k: the number of clusters
C: the original clustering solution that is obtained by applying k-means on X
For all clusters in clustering C do the following:
i = 1
$i < m$
No
Yes
Randomly replace 5% of X with noise points to obtain the new resampling data X^{new}
Apply k-means on X^{new} to obtain the clustering solution C^{new}
Calculate the Jaccard similarity between each cluster in the original clustering C and the most similar cluster in C^{new}
$j_{c,i} = \text{Max}\{\text{Jaccard}(c, d)\}$
where $d \in C^{new}$ and $c \in C$ and the noise points are excluded from C^{new} and C
i = i + 1
Calculate the stability score (stability) for each cluster c in clustering C
$stability\ (c) = \frac{1}{m}\sum_{i=1}^{m} F(j_{c,i})$
where $F(x)$: if$(x > 0.75)$ return(1), else return(0)
Calculate the stability score for clustering C by averaging the stability scores over its clusters $\{c_1, \ldots, c_k\}$
$\frac{1}{k}\sum_{i=1}^{k} stability\ (c)$
where $c \in C$
End

Figure 14: The resampling framework to compute the stability measure for each clustering. The input includes N data points $X = \{x_1, ..., x_N\}$, the number of clusters k , the number of resamplings m , and the clustering C that is obtained by applying k-means on X. This analysis generates m resampling data by randomly replacing 5% of data points with the noise, and computes m resampled clusterings based on k-means clustering. Each cluster c in C is compared with the most similar cluster in the resampling clustering, and the Jaccard coefficient between the two clusters is computed, while the noise points are excluded. The percentage of the times that Jaccard coefficients are larger than 0.75 is considered the stability measure for cluster c. The average of stability measures for all clusters belonging to clustering C is calculated and considered as the overall stability measure for clustering C.

the cluster c is compared to all the clusters in the clustering that is obtained from the resample data ($clustering_{i,j}$) based on the Jaccard distance. *The Jaccard coefficient [117], a similarity measure between sets, is used to compute the similarity between two clusters as follows:*

$$\mathrm{J(A,B)} = \frac{|A \cap B|}{|A \cup B|}, \quad A, B \subseteq X$$

where the term A and B are two clusters, consisting of some data points in $X = \{x_1, \cdots, x_N\}$.

If the Jaccard similarity between the cluster c (from the original clustering $clustering_i$) and the most similar cluster in the resampled clustering is equal or greater than 0.75, that cluster is considered stable (preserved). Thus, the stability of each cluster in $clustering_i$ is calculated as the percentage of the times that cluster is preserved (Jaccard coefficient ≥ 0.75) across the m different resamplings.

We then average the stability measures of the k clusters belonging to $clustering_i$, and consider it as the overall stability measure of $clustering_i$. Among n different clustering solutions ($clustering_i$ ($i \in [1..n]$)), we select the clustering solution with the maximum stability measure as the final clustering solution.

Figure 14 shows the detail of the resampling method we performed to compute the stability measure for each clustering. The clusters that are obtained by applying k-mean on the resampled dataset are compared with the clusters from the original input data only based on the non-noise points (the noise data points are excluded when two clusters are compared based on the Jaccard similarity metric.

3.4.5 Validation methods

We use 13 different datasets in which the cell types (labels) are known. To measure the level of similarity between the reference labels and the inferred labels that are obtained by each clustering method, we use three different metrics: adjusted rand index (ARI), adjusted mutual information (AMI), and V-measure as explained in the following.

Adjusted Rand Index

Given the cell labels, the Adjusted Rand Index (ARI) [113] is used to assess the similarity between the inferred clustering and the true clustering. ARI ranges from 0, for poor matching (a random clustering), to 1 for a perfect agreement with the true clustering. For a set of n data points, the contingency table is constructed based on the shared number of data points between two clusters. Suppose X and Y represent two different clusterings (representing the row and column of the contingency table, respectively) of n data points . X_i and Y_j denote a cluster in clusterings X and Y, and i and j refer to the row number and the column number of the contingency table, respectively. The ARI is defined as follow:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_i}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_i}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_i}{2}]/\binom{n}{2}} \tag{3.1}$$

where n_{ij} denotes the number of shared data points between clusters X_i and Y_j, and $a_i = \sum_k n_{ik}$ (the sum of the i^{th} row), and $b_j = \sum_k n_{kj}$ (the sum of the j^{th} column of the contingency table).

Adjusted Mutual Information

The adjusted mutual information (AMI) [274, 275] is a variation of mutual information that corrects for random partitioning, similar to the way the ARI corrects the rand index.

As explained in the previous section, given two different clusterings $X = \{X_1, X_2, ..., X_R\}$ and $Y = \{Y_1, Y_2, ..., Y_C\}$ of n data points with R and C clusters, respectively, the mutual information of cluster overlap between X and Y can be summarized as a contingency table $M_{R\times C} = [n_{ij}]$, where $i = 1...R$, $j = 1...C$, and n_{ij} represents the number of common data points between clusters X_i and Y_j. Suppose a data point is picked at random from X, the probability that the data point falls into cluster X_i is $p(i) = \frac{|X_i|}{n}$. The entropy [245] associated with the clustering X is calculated as follows:

$$H(X) = \sum_{i=1}^{R} P(i)\ logP(i) \tag{3.2}$$

$H(X)$ is non-negative and takes the value 0 only when there is no uncertainty determining a data point's cluster membership (there is only one cluster). The mutual information (MI) between two clusterings X and Y is calculated as follows:

$$MI(X,Y) = \sum_{i=1}^{R}\sum_{j=1}^{C} P(i,j)\ log\frac{P(i,j)}{P(i)P(j)} \tag{3.3}$$

where $P(i,j)$ denotes the probability that a data point belongs to both the cluster X_i in X and the cluster Y_j in Y:

$$P(i,j) = \frac{|X_i \cap Y_j|}{n} \tag{3.4}$$

MI is a non-negative quantity upper bounded by the entropies H(X) and H(Y). It quantifies the information shared by the two clusterings and therefore can be considered as a clustering similarity measure. The adjusted measure for the mutual information is defined as follows:

$$AMI(X,Y) = \frac{MI(X,Y) - E\{MI(X,Y)\}}{max\{H(X),H(Y)\} - E\{MI(X,Y)\}} \tag{3.5}$$

where the expected mutual information between two random clusterings is:

$$E\{MI(X,Y)\} = \sum_{i=1}^{R}\sum_{j=1}^{C} \sum_{n_{ij}=max(1,a_i+b_j-n)}^{min(a_i,b_j)} \frac{n_{ij}}{n} log(\frac{n.n_{ij}}{a_i b_j}) \frac{a_i!b_j!(n-a_i)!(n-b_j)!}{n!n_{ij}!(a_i-n_{ij})!(b_j-n_{ij})!(n-a_i-b_j+n_{ij})!} \tag{3.6}$$

where the a_i and b_j are the partial sums of the contingency table: $a_i = \sum_{j=1}^{C} n_{ij}$ and $b_j = \sum_{i=1}^{R} n_{ij}$.

The adjusted mutual information (AMI) takes a value of 1 when the two clusterings are identical and 0 when the MI between two partitions equals the value expected due to chance alone.

V-measure

The V-measure [226] is the harmonic mean between two measures: homogeneity and completeness. The homogeneity criteria is satisfied if a clustering assigns *only* those data points that are members of a single class (true cluster) to a single cluster. Thus, the class distribution within each cluster should be skewed to a single class (zero entropy) [226]. To determine how close a given clustering is to this ideal, the conditional entropy of the class distribution given the identified clustering is computed as $H(C \mid K)$, where $C = \{C_1, C_2, ..., C_l\}$ is a set of classes and K is a clustering $K = \{K_1, K_2, ..., K_m\}$. In the perfectly homogeneous case, this value is 0. However, this value is dependent on the size of the dataset and the distribution of class sizes [226]. Thus, this conditional entropy is normalized by the maximum reduction in entropy the clustering information could provide, $H(C)$. Therefore, the homogeneity is defined as follows:

$$h = \begin{cases} 1 & \text{if } H(C,K) = 0 \\ 1\text{-}\frac{H(C|K)}{H(C)} & \text{otherwise} \end{cases} \tag{3.7}$$

The completeness is symmetrical to homogeneity [226]. In order to satisfy the completeness criteria, a clustering must assign *all* of those data points that are members of a single class to a single cluster. To measure the completeness, the distribution of cluster assignments within each class is assessed [226]. In a perfectly complete clustering solution, each of these distributions will be completely skewed to a single cluster.

Given the homogeneity h and completeness c, the V-measure is computed as the weighted harmonic mean of homogeneity and completeness:

$$\text{V-measure} = \frac{(1+\beta) * h * c}{(\beta * h) + c} \tag{3.8}$$

if β is greater than 1, completeness is weighted more strongly in the calculation [226]. If β is less than 1, homogeneity is weighted more strongly. Since the computations of homogeneity, completeness and V-measure are completely independent of the number of classes, the number of clusters, the size of the dataset and the clustering algorithm, these measures can be employed for evaluating any clustering solution.

3.5 Results and validation

Tables 26–28 shows the comparison between the proposed method and five other methods: RaceID [93], SC3 [139], SEURAT [162], SINCERA [95], and SNN-Cliq [291] using the three metrics: ARI, AMI, and V-measures, respectively.

We used the R package fpc [102] to compute the k-means clustering based on the resampling method. We generated 20 different clusterings, and for each clustering we computed 1,000 clusterings based on the resampled datasets to find the most meaningful clustering. We used the log-transformation ($M' = log2(M + 1)$) for all methods except SINCERA. For SINCERA we followed the authors instructions [95] and used the original z-score normalization instead of the log-transformation. In order to generate SC3 results, we used the R package SC3 (http://bioconductor.org/packages/SC3, v.1.8.0). We applied the same gene filtering approach that authors proposed in their study (parameter gene_filter=TRUE).

For SEURAT we used the Seurat R package (v.2.3.4) [35]. We performed the t-SNE using the Rtsne R package with the default parameters, and we used DBSCAN algorithm for clustering. We ran SNN-cliq with the default parameters that are provided by the authors [291]. For RaceID, we used the R code provided by the authors [93] (https://github.com/dgrun/RaceID).

As shown in Figure 15, the proposed method performs better than the five methods across 13 different datasets. In this figure, the three boxplots shows the the performance of each method on these 13 datasets based on the adjusted rand index (ARI), adjusted mutual information (AMI), and V-measure. We performed the proposed method, SC3 and RaceID on each dataset for 50, 5, and 50 times, respectively. In these three methods, we calculated the average of ARIs, AMIs, and V-measures over different runs. Since SC3 is reported as a stable method by the authors [139], we run it only 5 times. Indeed, we have observed the results with a very small standard deviation in all 5 runs for all 13 datasets

Table 26: A comparison between the results of six methods: proposed, RaceID, SC3, Seurat, SINCERA, and SNN-Cliq. The adjusted rand index (ARI) [113] is used to evaluate the performance of each clustering method. The proposed method, RaceID, and SC3 are performed 50, 50, and 5 times on each dataset, respectively. SC3 was performed only 5 times because it is very stable (standard deviation of zero for all datasets). The average ARIs across different runs are computed for the proposed method, SC3, and RaceID. Since SNN-Cliq, SINCERA and SEURAT are deterministic, they are performed only once. The proposed method was the best for 8 out of the 13 datasets. The proposed method also yielded the best average ARI, as shown in Figure 15.

Dataset	#cell types	Proposed		RaceID		SC3	
		K (mean±sd)	ARI (mean)	K (mean±sd)	ARI (mean)	K (mean±sd)	ARI (mean±sd)
Biase	3	3±0	0.94±0.01	3.14±0.6	0.84±0.25	3±0	0.94±0
Deng	10	10±0	0.58±0.02	1	0±0	9±0	0.65±0
Goolam	5	3±0	0.8±0.09	1	0±0	6±0	0.59±0
Klein	4	6±0	0.69±0.01	2.98±0.14	0.48±0.001	19±0	0.44±0.01
Patel	5	5±0	0.66±0.09	7.44±1.88	0.66±0.08	17±0	0.45±0.01
Pollen	11	8±0	0.86±0.02	8.36±2.27	0.55±0.11	10±0	0.93±0
Treutlein	5	3±0	0.72±0.03	1±0	0±0	3±0	0.66±0
Yan	8	5±0	0.81±0.02	5.5±2.34	0.55±0.17	4±0	0.76±0
sim3	3	3±0	1±0±0	1±0	0±0	3±0	1±0
sim4	4	4±0	0.99±0.005	1±0	0±0	4±0	0.99±0.0005
sim6	6	7.9±0.3	0.56±0.03	1±0	0±0	3±0	0.53±0
sim8	8	9.34±0.47	0.77±0.03	1±0	0±0	4±0	0.53±0.04
sim±Tung	8	8±0	0.42±0	1±0	0±0	8±0	0±0

Dataset	#cell types	SINCERA		SNN-Cliq		Seurat	
		K	ARI	K	ARI	K	ARI
Biase	3	6	0.71	6	0.66	4	0.78
Deng	10	3	0.42	17	0.4	6	0.45
Goolam	5	13	0.19	17	0.2	3	0.05
Klein	4	43	0.45	265	0.11	3	0
Patel	5	10	0.78	26	0.14	5	0.63
Pollen	11	10	0.9	22	0.71	8	0.85
Treutlein	5	7	0.35	5	0.62	1	0
Yan	8	8	0.59	13	0.79	3	0.56
sim3	3	120	0.12	147	0.03	3	1
sim4	4	464	0.08	437	0.01	3	0.57
sim6	6	68	0.25	143	0.06	6	1
sim8	8	68	0.35	290	0.05	8	1
sim_Tung	8	17	0.001	77	0.001	8	0

confirming the claims of the authors. The other clustering methods SEURAT, SINCERA, and SNN-Cliq were run only once since they are deterministic.

3.6 Discussion

The results shown in Tables 26–28 merit some discussion. The Goolam dataset, for instance, includes 5 true cell types. On this dataset, the proposed algorithm identifies 3 clusters, while SC3 identifies 6, RaceID 1, Seurat 2, SINCERA 13 and SNN-Cliq 17 types. Even though the number of clusters closest to the number of true types is 6, as yielded by

Table 27: A comparison between the results of six methods: proposed, RaceID, SC3, Seurat, SINCERA, and SNN-Cliq.The adjusted mutual information (AMI) [274, 275], is used to evaluate the performance of each clustering method. The proposed method, RaceID, and SC3 are performed 50, 50, and 5 times on each dataset, respectively. The average AMIs across different runs are computed for the proposed method, SC3, and RaceID. Since SNN-Cliq, SINCERA and SEURAT are deterministic, they are performed only once.

Dataset	#cell types	Proposed		RaceID		SC3	
		K (mean±sd)	AMI (mean±sd)	K (mean±sd)	K (mean±sd)	K (mean±sd)	AMI (mean±sd)
Biase	3	3±0	0.92±0.02	3.14±0.6	0.85±0.23	3±0	0.92±0
Deng	10	10±0	0.73±0.01	1±0	0±0	9±0	0.81±0
Goolam	5	3±0	0.73±0.04	1±0	0±0	6±0	0.69±0
Klein	4	6±0	0.67±0.06	2.98±0.14	0.51±0.05	19±0	0.53±0
Patel	5	5±0	0.86±0.01	7.44±1.88	0.66±0.1	17±0	0.93±0
Pollen	11	8±0	0.72±0.01	8.36±2.27	0.68±0	10±0	0.53±0.01
Treutlein	5	3±0	0.54±0.03	1±0	0±0	3±0	0.62±0
Yan	8	5±0	0.78±0.01	5.5±2.34	0.61±0.17	4±0	0.72±0
sim3	3	3±0	1±0	1±0	0±0	3±0	1±0
sim4	4	4±0	0.99±0.007	1±0	0±0	4±0	0.99±0.001
sim6	6	7.9±0.3	0.64±0.02	1±0	0±0	3±0	0.51±0
sim8	8	9.34±0.47	0.85±0.01	1±0	0±0	4±0	0.56±0
sim_Tung	8	8±0	0.51±0.008	1±0	0±0	8±0	0.006±0

Dataset	#cell types	SINCERA		SNN-Cliq		Seurat	
		K	AMI	K	AMI	K	AMI
Biase	3	6	0.64	6	0.62	4	0.74
Deng	10	3	0.48	17	0.6	6	0.59
Goolam	5	13	0.4	17	0.42	3	0.11
Klein	4	43	0.52	265	0.21	3	0.06
Patel	5	10	0.73	26	0.31	5	0.68
Pollen	11	10	0.91	22	0.74	8	0.87
Treutlein	5	7	0.46	5	0.51	1	0
Yan	8	8	0.72	13	0.76	3	0.58
sim3	3	120	0.23	147	0.21	3	1
sim4	4	464	0.21	437	0.2	3	0.66
sim6	6	68	0.42	143	0.3	6	1
sim8	8	68	0.51	290	0.31	8	1
sim_Tung	8	17	0.04	77	0.13	8	0

SC3, the membership of various cells in these clusters is not correct since the ARI index associated to these 6 clusters is only 0.59 compared to the ARI index of 0.8 associated to the 3 clusters constructed by the proposed method.

Conversely, for the Patel dataset that includes 5 cell types, the proposed method was able to correctly estimate the number of clusters (k=5). However, the distribution of the individual cells across these five clusters is not perfect, as illustrated by the lower ARI value of 0.66, compared to the 0.78 ARI associated with the SINCERA results.

As another observation, the Pollen dataset includes 11 cell types. Using this dataset, the number of clusters (k=10) determined by SINCERA is close to the correct number of cell

Table 28: A comparison between the results of six methods: proposed, RaceID, SC3, Seurat, SINCERA, and SNN-Cliq. The V-measure [226] is used to evaluate the performance of each clustering method. The proposed method, RaceID, and SC3 are performed 50, 50, and 5 times on each dataset, respectively. The average V-measures across different runs are computed for the proposed method, SC3, and RaceID. Since SNN-Cliq, SINCERA and SEURAT are deterministic, they are performed only once.

Dataset	#cell types	Proposed		RaceID		SC3	
		K (mean±sd)	V-measure (mean)	K (mean±sd)	V-measure (mean)	K (mean±sd)	V-measure (mean)
Biase	3	3±0	0.93±0.03	3.14±0.6	0.87±0.2	3	0.93±0
Deng	10	10±0	0.72±0.01	1±0	0±0	9	0.74±0.001
Goolam	5	3±0	0.82±0.04	1±0	0±0	6	0.98±0
Klein	4	6±0	0.38±0.01	2.98±0.14	0.4±0.06	19	0.31±0.002
Patel	5	5±0	0.56±0.02	7.44±1.88	0.54±0.04	17	0.46±0.002
Pollen	11	8±0	0.95±0.01	8.36±2.27	0.76±0.03	10	0.93±0
Treutlein	5	3±0	0.96±0	1±0	0±0	3	0.89±0
Yan	8	5±0	0.83±0.02	5.5±2.34	0.68±0.07	4	0.81±0
sim3	3	3±0	1±0	1±0	0±0	3	1±0
sim4	4	4±0	0.99±0.0002	1±0	0±0	4	0.99±0.00003
sim6	6	7.9±0.3	0.98±0	1±0	0±0	3	0.97±0.0004
sim8	8	9.34±0.47	0.99±0	1±0	0±0	4	0.98±0.004
sim_Tung	8	8±0	0.96±0.03	1±0	0±0	8	0.66±0

Dataset	#cell types	SINCERA		SNN-Cliq		Seurat	
		K	V-measure	K	V-measure	K	V-measure
Biase	3	6	0.72	6	0.7	4	0.73
Deng	10	3	0.93	17	0.64	6	0.93
Goolam	5	13	0.71	17	0.65	3	0.66
Klein	4	43	0.36	265	0.29	3	0.46
Patel	5	10	0.55	26	0.44	5	0.62
Pollen	11	10	0.94	22	0.72	8	0.93
Treutlein	5	7	0.93	5	0.92	1	0
Yan	8	8	0.65	13	0.78	3	0.73
sim3	3	120	0.95	147	0.95	3	1
sim4	4	464	0.97	437	0.97	3	0.96
sim6	6	68	0.97	143	0.97	6	1
sim8	8	68	0.98	290	0.98	8	1
sim_Tung	8	17	0.82	77	0.8	8	0.66

types. However, SC3 achieved better clustering (ARI=0.93) in contrast to the five other methods. SC3 identified 17 different clusters using this dataset.

Two conclusions may be drawn from these observations. First, results should not be assessed based on the agreement between the number of clusters found and the number of known cell types – the assignment of each cell to a given type is more important. Second, larger number of clusters reported will be associated with larger values of ARI. Therefore, results that include very large number of clusters should be regarded with caution.

RaceID and Seurat both were not able to find a meaningful clustering for the Treutlein dataset. The identified number of clusters by both RaceID and Seurat is 1 (k=1), while this dataset includes 5 different cell types. As a result, the clusterings obtained by these

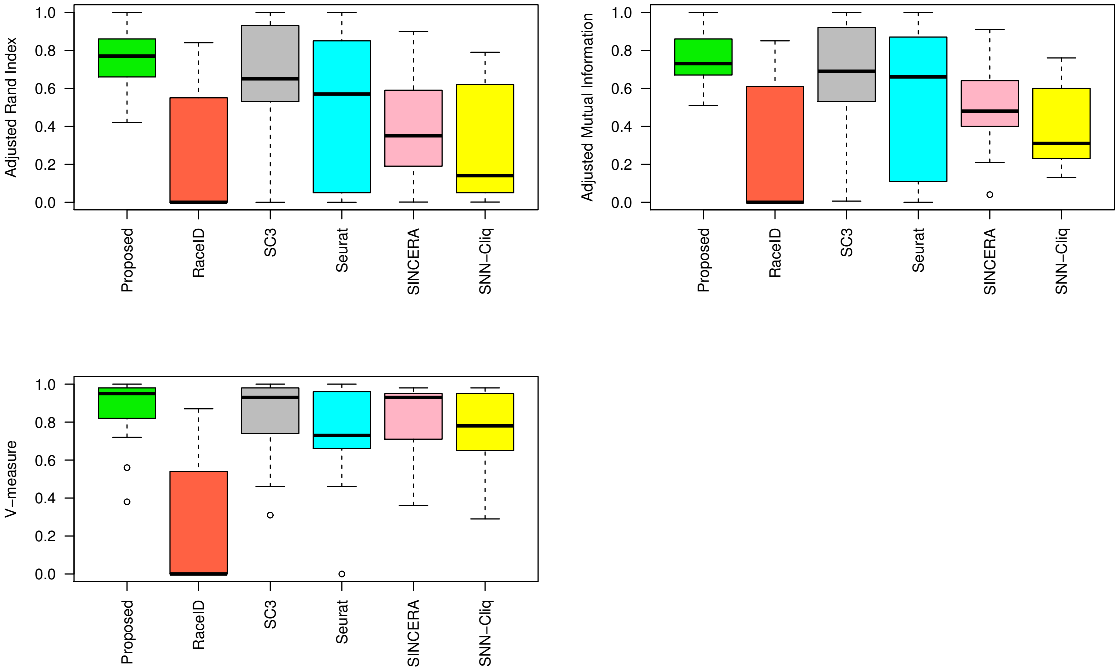


Figure 15: The performance comparison using 13 single cell datasets based on three metrics: the adjusted rand index (ARI), adjusted mutual information (AMI), and V-measure. The proposed method and RaceID were applied 50 times on each dataset. SC3 was used only 5 times on each dataset because it is very stable. The average ARIs, AMIs, and V-measures across different runs are computed for the proposed method, RaceID, and SC3. Since SNN-Cliq, SINCERA, and SEURAT are deterministic, they are run only once for each dataset.

two methods are poorly matched to the reference clustering. In Deng dataset, the best ARI of 0.65 is obtained by SC3 but this value is not very high. The poor results obtained by all 6 methods using this dataset might be due to noisy data.

We also assessed the reproducibility/stability of the stochastic methods: proposed [206], RaceID, and SC3 by running each method several times. Although SC3's consensus pipeline provides a very stable solution (very low standard deviation for the three metrics and k across all datasets), it is computationally more costly than other methods. In summary, one key advantage of our proposed method is that we produce consistent clustering across

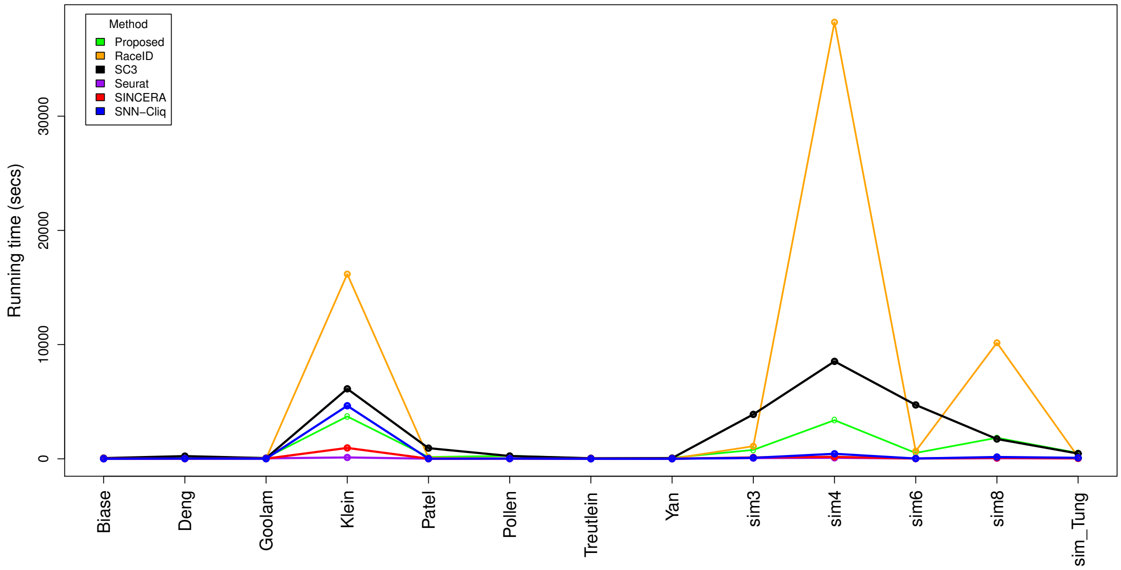


Figure 16: The run time of the different methods using 13 single cell datasets.

different datasets.

The run time for each method using 13 different datasets is shown in Figure 16. It is notable that RaceID, the proposed method, and SC3 have a non-linear increase in run time. At this time, it appears that it is unfeasible to perform this method on large datasets consisting of thousands of cells. The fastest method among all the methods is Seurat, which is a graph-based method. The graph-based methods often return only a single clustering solution with a faster run time and they do not require the user to provide the number of clusters [138]. Seurat is a popular choice for the large data sets based on the its optimal speed and scalability. However, it has been shown that Seurat does not provide an accurate solution for smaller datasets [138]. The details of the run times are shown in Table 29.

More generally, finding an optimal clustering method that provides stable solutions for all situations may not be possible. In fact, because no method can perform well for all

Table 29: The run time (secs) of the different methods using 13 single cell datasets, including 5 simulation datasets.

	Proposed	RaceID	SC3	Seurat	SINCERA	SNN-Cliq
Biase	43.47	8.79	59.79	27.3	3.56	0.62
Deng	188.51	55.37	234.3	23.68	10	16.39
Goolam	86.55	15.69	59.52	41.36	14.16	5.75
Klein	3709.28	16175.58	6127.56	117.1	952.25	4643.8
patel	154.18	100.32	933.05	13.78	10.25	12.22
Pollen	257.14	50.5	245.02	25.86	13.11	17.35
treutlein	53.49	8.34	47.89	23.09	5.82	1.33
Yan	87.74	13.46	60.31	19.5	4.96	2.23
sim3	777.08	1102.2	3888	63	118.8	84.71
sim4	3381.1	38232	8532	85.2	170.4	439.12
sim6	510.48	670.2	4716	31.09	18.3	23.01
sim8	1839.75	10160.64	1733.4	53.33	85.2	157.47
sim_Tung	504.6	171.6	458.4	40.66	37.36	86.13

situations, a comparative analysis of methods based on a set of criteria should be employed [138].

In the following, we investigate the application of the uniform manifold approximation and projection (UMAP) for reducing the dimensionality of the single cell data that is used for the resampling-based clustering framework. We also provide the performance of the clustering results using different noise tuning hyperparmaters.

3.6.1 Uniform manifold approximation and projection

We investigated the performance of the resampling-based k-means clustering method using two dimensionality reduction techniques: the t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP).

The UMAP is an algorithm for reducing the dimensionality of the data based on manifold learning techniques and the topological data analysis. Unlike t-SNE [161] that preserves only local structure in the data, UMAP [174] claims to preserve both local and the global structure in the data. The main difference between t-SNE and UMAP is the interpretation of the distance between objects or clusters [174].

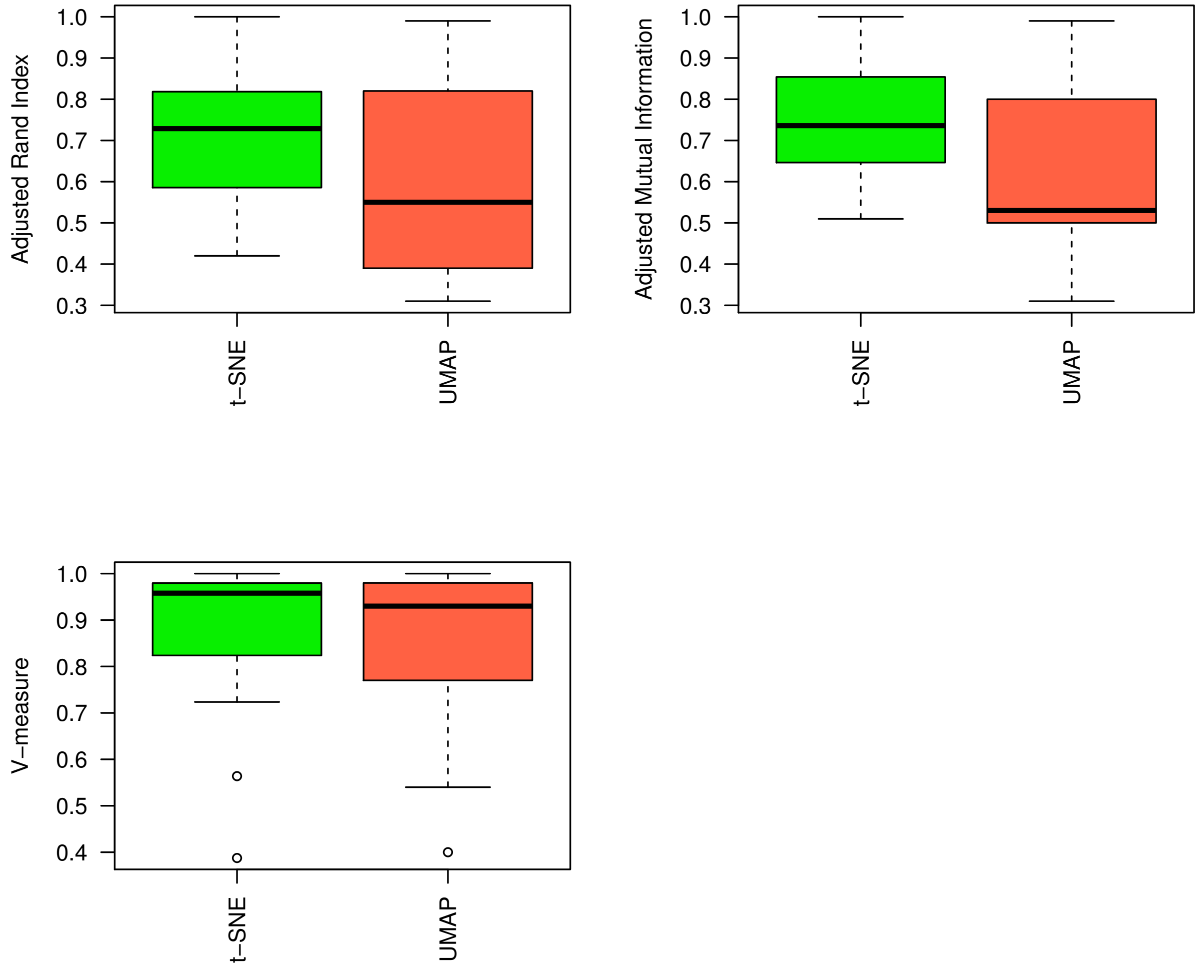


Figure 17: In this comparison, 13 datasets, including 8 real single cell gene expression datasets are used. The results are compared based on the three metrics: the adjusted rand index (ARI), adjusted mutual information (AMI), and V-measure. The results show that the t-SNE dimensionality reduction technique provides better performance in comparison to the UMAP technique.

We applied the UMAP [174] to reduce the dimensionality of the distance matrix in which the rows and columns are the cells. Then, we identified the most stable clustering of cells using the lower-dimension distance matrix based on the resampled-based k-means clustering. As shown in Figure 17, we found that the performance is better when t-SNE is used for the dimensionality reduction.

Tables 30–32 show the comparison between the results of six methods: K-means-UMAP, RaceID, SC3, Seurat, SINCERA, and SNN-Cliq .The K-means-UMAP is an alternative to the proposed method in which UMAP technique is employed for the dimensionality reduction.

Table 30: A comparison between the results of six methods: K-means-UMAP, RaceID, SC3, Seurat, SINCERA, and SNN-Cliq based on the adjusted rand index (ARI). The K-means-UMAP is an alternative to the proposed method in which UMAP technique is employed for the dimensionality reduction. The K-means-UMAP method, RaceID, and SC3 are performed 50, 50, and 5 times on each dataset, respectively. SC3 was performed only 5 times because it is very stable (standard deviation of zero for all datasets). The average ARIs across different runs are computed for the K-means-UMAP, SC3, and RaceID. Since SNN-Cliq, SINCERA and SEURAT are deterministic, they are performed only once. For each dataset, the best ARI is highlighted in green.

Dataset	#cell types	K-means-UMAP		RaceID		SC3		SINCERA		SNN-Cliq		Seurat	
		K (mean±sd)	ARI (mean±sd)	K (mean±sd)	ARI (mean±sd)	K (mean±sd)	ARI (mean±sd)	K	ARI	K	ARI	K	ARI
Biase	3	3±0	0.95±0	3.14±0.6	0.84±0.25	3±0	0.94±0	6	0.71	6	0.66	4	0.78
Deng	10	5±0	0.39±0.02	1±0	0±0	9±0	0.65±0.002	3	0.42	17	0.4	6	0.45
Goolam	5	2.88±0.33	0.82±0.11	1±0	0±0	6±0	0.59±0	13	0.19	17	0.2	3	0.05
Klein	4	3±0	0.55±0	2.98±0.14	0.48±0.001	19±0	0.44±0.01	43	0.45	265	0.11	3	0
Patel	5	2.76±0.52	0.31±0.07	7.44±1.88	0.66±0.08	17±0	0.45±0.01	10	0.78	26	0.14	5	0.63
Pollen	11	7.98±0.14	0.84±0.03	8.36±2.27	0.55±0.11	10±0	0.93±0	10	0.9	22	0.71	8	0.85
Treutlein	5	3.24±0.43	0.51±0.12	1±0	0±0	3±0	0.66±0	7	0.35	5	0.62	1	0
Yan	8	8±0	0.75±0.1	5.5±2.34	0.55±0.17	4±0	0.76±0	8	0.59	13	0.79	3	0.56
sim3	3	3±0	0.99±0.01	1±0	0±0	3±0	1±0	120	0.12	147	0.03	3	1
sim4	4	3±0	0.55±0	1±0	0±0	4±0	0.99±0.0005	464	0.08	437	0.01	3	0.57
sim6	6	16.72±0.45	0.32±0.02	1±0	0±0	3±0	0.53±0.005	68	0.25	143	0.06	6	1
sim8	8	18±0	0.38±0.02	1±0	0±0	4±0	0.53±0.04	68	0.35	290	0.05	8	1
sim_Tung	8	8±0	0.41±0.01	1±0	0±0	8±0	0±0	17	0.001	77	0.001	8	0

Table 31: A comparison between the results of six methods: K-means-UMAP, RaceID, SC3, Seurat, SINCERA, and SNN-Cliq based on the adjusted mutual information (AMI). The K-means-UMAP is an alternative to the proposed method in which UMAP technique is employed for the dimensionality reduction. The K-means-UMAP, RaceID, and SC3 are performed 50, 50, and 5 times on each dataset, respectively. The average AMIs across different runs are computed for the K-means-UMAP, SC3, and RaceID. Since SNN-Cliq, SINCERA and SEURAT are deterministic, they are performed only once.

Dataset	#cell types	K-means-UMAP		RaceID		SC3		SINCERA		SNN-Cliq		Seurat	
		K (mean±sd)	AMI (mean±sd)	K (mean±sd)	AMI (mean±sd)	K (mean±sd)	AMI (mean±sd)	K	AMI	K	AMI	K	AMI
Biase	3	3±0	0.92±0	3.14±0.6	0.85±0.23	3±0	0.92±0	6	0.64	6	0.62	4	0.74
Deng	10	5±0	0.51±0.01	1±0	0±0	9±0	0.81±0.006	3	0.48	17	0.6	6	0.59
Goolam	5	2.88±0.33	0.72±0.1	1±0	0±0	6±0	0.69±0	13	0.4	17	0.42	3	0.11
Klein	4	3±0	0.53±0	2.98±0.14	0.51±0.05	19±0	0.53±0.006	43	0.52	265	0.21	3	0.06
Patel	5	2.76±0.52	0.31±0.07	7.44±1.88	0.66±0.1	17±0	0.93±0	10	0.73	26	0.31	5	0.68
Pollen	11	7.98±0.14	0.85±0.02	8.36±2.27	0.68±0	10±0	0.53±0.01	10	0.91	22	0.74	8	0.87
Treutlein	5	3.24±0.43	0.44±0.06	1±0	0±0	3±0	0.62±0	7	0.46	5	0.51	1	0
Yan	8	8±0	0.80±0.04	5.5±2.34	0.61±0.17	4±0	0.72±0	8	0.72	13	0.76	3	0.58
sim3	3	3±0	0.99±0.02	1±0	0±0	3±0	1±0	120	0.23	147	0.21	3	1
sim4	4	4±0	0.61±0	1±0	0±0	4±0	0.99±0.001	464	0.21	437	0.2	3	0.66
sim6	6	16.72±0.45	0.48±0.01	1±0	0±0	3±0	0.51±0.004	68	0.42	143	0.3	6	1
sim8	8	18±0	0.53±0.01	1±0	0±0	4±0	0.56±0.007	68	0.51	290	0.31	8	1
sim_Tung	8	8±0	0.50±0.01	1±0	0±0	8±0	0.006±0	17	0.04	77	0.13	8	0

Table 32: A comparison between the results of six methods: K-means-UMAP, RaceID, SC3, Seurat, SINCERA, and SNN-Cliq based on the V-measure. The K-means-UMAP is an alternative to the proposed method in which UMAP technique is employed for the dimensionality reduction. The K-means-UMAP, RaceID, and SC3 are performed 50, 50, and 5 times on each dataset, respectively. The average V-measures across different runs are computed for the K-means-UMAP, SC3, and RaceID. Since SNN-Cliq, SINCERA and SEURAT are deterministic, they are performed only once. For each dataset, the best V-measure is highlighted in green.

Dataset	#cell types	K-means-UMAP		RaceID		SC3		SINCERA		SNN-Cliq		Seurat	
		K (mean±sd)	V-measure (mean±sd)	K (mean±sd)	V-measure (mean±sd)	K (mean±sd)	V-measure (mean±sd)	K	V-measure	K	V-measure	K	V-measure
Biase	3	3±0	0.93±0	3.14±0.6	0.87±0.2	3	0.93±0	6	0.72	6	0.7	4	0.73
Deng	10	5±0	0.75±0.06	1±0	0±0	9	0.74±0.001	3	0.93	17	0.64	6	0.93
Goolam	5	2.88±0.33	0.84±0.04	1±0	0±0	6	0.98±0	13	0.71	17	0.65	3	0.66
Klein	4	3±0	0.40±0	2.98±0.14	0.4±0.06	19	0.31±0.002	43	0.36	265	0.29	3	0.46
Patel	5	2.76±0.52	0.54±0.02	7.44±1.88	0.54±0.04	17	0.46±0.002	10	0.55	26	0.44	5	0.62
Pollen	11	7.98±0.14	0.94±0.02	8.36±2.27	0.76±0.03	10	0.93±0	10	0.94	22	0.72	8	0.93
Treutlein	5	3.24±0.43	0.94±0.01	1±0	0±0	3	0.89±0	7	0.93	5	0.92	1	0
Yan	8	8±0	0.77±0.07	5.5±2.34	0.68±0.07	4	0.81±0	8	0.65	13	0.78	3	0.73
sim3	3	3±0	1±0	1±0	0±0	3	1±0	120	0.95	147	0.95	3	1
sim4	4	3±0	0.98±0	1±0	0±0	4	0.99±0.00003	464	0.97	437	0.97	3	0.96
sim6	6	16.72±0.45	0.98±0	1±0	0±0	3	0.97±0.0004	68	0.97	143	0.97	6	1
sim8	8	18±0	0.99±0	1±0	0±0	4	0.98±0.004	68	0.98	290	0.98	8	1
sim_Tung	8	8±0	0.91±0.01	1±0	0±0	8	0.66±0	17	0.82	77	0.80	8	0.66

3.6.2 Noise tuning for the resampling-based clustering approach

We replaced the 5% of the data points with noise (noise tuning threshold=0.05). In [100], Hennig performed a comparative analysis using two thresholds 0.05 and 0.2 and showed that the threshold 0.05 provides more stable clusterings. We assessed the performance of our proposed method using other noise tuning thresholds: 0.1 and 0.2. Tables 33–35 show comparison between the proposed method (using 3 thresholds) and five other methods based on the three metrics: the adjusted rand index (ARI), adjusted mutual information (AMI), and V-measure. Indeed, we have observed the threshold 0.05 provides better performance.

To assess the stability of a cluster, we used the same threshold (0.75) that is recommended by Hennig [100]. In this study [100], it has been shown that a stable cluster will yield a Jaccard similarity value of 0.75 or more. Thus, if the Jaccard similarity between the cluster (from original clustering) and the most similar cluster in the resampled clustering is equal or greater than 0.75, that the cluster is considered as successfully recovered.

Table 33: A comparison between the results of six methods: proposed (noise tuning thresholds: 0.05, 0.1 and 0.2), RaceID, SC3, Seurat, SINCERA, and SNN-Cliq. The adjusted rand index (ARI) [113] is used to evaluate the performance of each clustering method. The proposed method was performed based on three noise tuning thresholds: 0.05, 0.1 and 0.2. The proposed method, RaceID, and SC3 are performed 50, 50, and 5 times on each dataset, respectively. SC3 was performed only 5 times because it is very stable (standard deviation of zero for all datasets). The average ARIs across different runs are computed for the proposed method, SC3, and RaceID. Since SNN-Cliq, SINCERA and SEURAT are deterministic, they are performed only once. For each dataset, the best ARI is highlighted in green.

Dataset	#cell types	Proposed- ARI (mean±sd)			RaceID	SC3	SINCERA	SNN-Cliq	Seurat
		Threshold=0.05	Threshold=0.1	Threshold=0.2	ARI (mean±sd)	ARI (mean±sd)	ARI	ARI	ARI
Biase	3	0.94±0.01	0.95±0.01	0.91±0.15	0.84±0.25	0.94±0	0.71	0.66	0.78
Deng	10	0.58±0.02	0.43±0.03	0.43±0.01	9±0	0.65±0.002	0.42	0.4	0.45
Goolam	5	0.80±0.09	0.75±0.12	0.69±0.13	0±0	0.59±0	0.19	0.20	0.05
Klein	4	0.69±0.01	0.70±0.01	0.70±0.03	0.48±0.001	0.44±0.01	0.45	0.11	0
Patel	5	0.66±0.09	0.96±0.01	0.95±0.01	0.66±0.08	0.45±0.01	0.78	0.14	0.63
Pollen	11	0.86±0.02	0.89±0.04	0.85±0.05	0.55±0.11	0.93±0	0.90	0.71	0.85
Treutlein	5	0.72±0.03	0.32±0.04	0.31±0.04	0±0	0.66±0	0.35	0.62	0
Yan	8	0.81±0.02	0.82±0.09	0.90±0.03	0.55±0.17	0.76±0	0.59	0.79	0.56
sim3	3	1±0	1±0	1±0	0±0	1±0	0.12	0.03	1
sim4	4	0.99±0.005	0.90±0.08	0.89±0.08	0±0	0.99±0.0005	0.08	0.01	0.57
sim6	6	0.56±0.03	0.59±0.03	0.56±0.05	0±0	0.53±0.005	0.25	0.06	1
sim8	8	0.77±0.03	0.78±0.04	0.79±0.045	0±0	0.53±0.04	0.35	0.05	1
sim_Tung	8	0.42±0	0±0	0±0	0±0	0±0	0.001	0.001	0

Table 34: A comparison between the results of six methods: proposed (noise tuning thresholds: 0.05, 0.1 and 0.2), RaceID, SC3, Seurat, SINCERA, and SNN-Cliq. The adjusted mutual information (AMI) [274, 275] is used to evaluate the performance of each clustering method. The proposed method was performed based on three noise tuning thresholds: 0.05, 0.1 and 0.2. The proposed method, RaceID, and SC3 are performed 50, 50, and 5 times on each dataset, respectively. The average AMIs across different runs are computed for the proposed method, SC3, and RaceID. Since SNN-Cliq, SINCERA and SEURAT are deterministic, they are performed only once. For each dataset, the best AMI is highlighted in green.

Dataset	#cell types	Proposed- AMI (mean±sd)			RaceID	SC3	SINCERA	SNN-Cliq	Seurat
		Threshold=0.05	Threshold=0.1	Threshold=0.2	AMI (mean±sd)	AMI (mean±sd)	AMI	AMI	AMI
Biase	3	0.92±0.02	0.88±0.15	0.90±0.07	0.85±0.23	0.92±0	0.64	0.62	0.74
Deng	10	0.73±0.01	0.58±0.03	0.58±0.02	0±0	0.81±0.006	0.48	0.6	0.59
Goolam	5	0.73±0.04	0.73±0.05	0.71±0.05	0±0	0.69±0	0.4	0.42	0.11
Klein	4	0.67±0.06	0.73±0.01	0.73±0.02	0.51±0.05	0.53±0.006	0.52	0.21	0.06
Patel	5	0.86±0.01	0.94±0.03	0.94±0.01	0.66±0.1	0.93±0 0	0.73	0.31	0.68
Pollen	11	0.72±0.01	0.89±0.02	0.88±0.02	0.68±0	0.53±0.01	0.91	0.74	0.87
Treutlein	5	0.54±0.03	0.41±0.04	0.42±0.04	0±0	0.62±0	0.46	0.51	0
Yan	8	0.78±0.01	0.78±0.01	0.89±0.04	0.61±0.17	0.72±0	0.72	0.76	0.58
sim3	3	1±0	1±0	1±0.01	0±0	1±0	0.23	0.21	1
sim4	4	0.99±0.007	0.91±0.07	0.89±0.07	0±0	0.99±0.001	0.21	0.2	0.66
sim6	6	0.64±0.02	0.66±0.02	0.65±0.03	0±0	0.51±0.004	0.42	0.3	1
sim8	8	0.85±0.01	0.85±0.02	0.85±0.02	0±0	0.56±0.007	0.51	0.31	1
sim_Tung	8	0.51±0.008	0.01±0	0.01±0	0±0	0.006±0	0.04	0.13	0

Table 35: A comparison between the results of six methods: proposed (noise tuning thresholds: 0.05, 0.1 and 0.2), RaceID, SC3, Seurat, SINCERA, and SNN-Cliq. The V-measure [226] is used to evaluate the performance of each clustering method. The proposed method was performed based on three noise tuning thresholds: 0.05, 0.1 and 0.2. The proposed method, RaceID, and SC3 are performed 50, 50, and 5 times on each dataset, respectively. The average V-measures across different runs are computed for the proposed method, SC3, and RaceID. Since SNN-Cliq, SINCERA and SEURAT are deterministic, they are performed only once. For each dataset, the best V-measure is highlighted in green.

Dataset	#cell types	Proposed- V-measure (mean±sd)			RaceID	SC3	SINCERA	SNN-Cliq	Seurat
		Threshold=0.05	Threshold=0.1	Threshold=0.2	V-measure (mean±sd)	V-measure (mean±sd)	V-measure	V-measure	V-measure
Biase	3	0.93±0.03	0.93±0.02	0.90±0.1	0.87±0.2	0.93±0	0.72 6	0.7	0.73
Deng	10	0.72±0.01	0.75±0.06	0.81±0.1	0±0	0.74±0.001	0.93	0.64	0.93
Goolam	5	0.82±0.04	0.86±0.09	0.85±0.09	0±0	0.98±0	0.71	0.65	0.66
Klein	4	0.38±0.01	0.39±0.01	0.40±0.03	0.40±0.02	0.31±0.002	0.36	0.29	0.46
Patel	5	0.56±0.02	0.81±0.03	0.81±0.03	0.54±0.04	0.46±0.002	0.55	0.44	0.62
Pollen	11	0.95±0.01	0.92±0.02	0.91±0.02	0.76±0.03	0.93±0	0.94	0.72	0.93
Treutlein	5	0.96±0	0.93±0.01	0.93±0.01	0±0	0.89±0	0.93	0.92	0
Yan	8	0.83±0.02	0.85±0.04	0.87±0.04	0.68±0.07	0.81±0	0.65	0.78	0.73
sim3	3	1±0	1±0	1±0	0±0	1±0	0.95	0.95	1
sim4	4	0.99±0.0002	0.99±0.003	0.99±0	0±0	0.99±0.00003	0.97	0.97	0.96
sim6	6	0.98±0	0.99±0.001	0.99±0	0.99±0	0.99±0.0004	0.97	0.97	1
sim8	8	0.99±0	0.99±0.0009	1±0	0±0	0.98±0.004	0.98	0.98	1
sim_Tung	8	0.96±0.03	0.71±0.01	0.71±0.01	0±0	0.66±0	0.82	0.80	0.66

3.7 Summary

Recent advances in single-cell RNA-Seq (scRNASeq) provide the opportunity to perform single-cell transcriptome analysis. In this thesis, we develop a pipeline to cluster the individual cells based on their gene expression values such that each cluster consisting of cells with specific functions or distinct developmental stages. We first filter genes that are not expressed in any cell. Then, we compute the distance between the cells using the Euclidean distance. We reduce the dimensions of the distance matrix data using the t-distributed stochastic neighbor embedding (t-SNE) technique. Based on the dimensionality reduced distance matrix, we explore strong patterns (clusters) of cells by randomly drawing a percentage of the data points without replacement, and replacing them with points from a noise distribution. We apply the proposed method on 13 different single cell datasets, and we compare it with five related methods: RaceID, SC3, Seurat, SINCERA, and SNN-Cliq. The results of the evaluation on datasets demonstrate that the proposed method yields better clustering results in comparison to the existing methods.

CHAPTER 4 FUTURE WORK

In the first part of this thesis, we proposed a novel approach that discovers candidate drugs for repurposing. This approach first builds a drug-disease network by considering all interactions between drug targets and disease-related genes in the context of signaling pathways. This network is integrated with gene-expression measurements to identify drugs with new desired therapeutic effects. The results are assessed using a true gold standard, that of the already FDA-approved drugs. The already FDA-approved drugs for a given condition are the gold standard because such drugs successfully passed all the preclinical and clinical trials for that condition and were demonstrated to be efficacious in each condition. We provided evidences that support the usefulness of the proposed candidates in treatment of the human diseases: idiopathic pulmonary fibrosis (IPF), non-small cell lung cancer, prostate cancer, breast cancer. In all diseases, the proposed approach is able to rank highly the approved drugs for the given conditions. For many of the proposed repurposing candidates, there is significant preliminary evidence, as well as a number of clinical trials in progress. Although our proposed framework is studied in the context of drug repurposing, it also can be used to identify novel targets for FDA-approved drugs and predicting their mechanism of action using the drug-disease networks constructed by the proposed approach.

Second, we proposed a resampling-based clustering approach that solves one of the most important challenges in the personalized medicine area, which is identification of cell types from single cell data. The proposed pipeline can be used to analyze and understand cellular heterogeneity and how this contributes to the biological system.

In particular, the proposed clustering framework can be used to analyze the impact of glaucoma on single retinal ganglion cells (RGCs) and study the RGC subtypes that are more resilient and susceptible to glaucoma insult [262, 302].

As future work, one could identify driver genes (cell type signatures) that vary between two or more clusters (cell types). The driver genes could be identified as genes that are highly expressed in only one of the clusters and are able to distinguish one cluster from the remaining clusters. Subsequently, one could use the impact analysis method [72] to perform gene and pathway enrichment analysis using the discovered driver genes.

Next, we are interested in upgrading the proposed clustering framework to identify the rare cell populations in data (e.g cancer stem cells). We will generate synthetic data with rare cell types to evaluate the sensitivity our framework for rare cell type identification.

The proposed clustering framework on other data types (e.g mRNA, miRNA, and methylation data) can be used to distinguish between subgroups of patients (respondent vs. non-respondents) as well as disease subtypes (aggressive vs. non-aggressive). In addition to subtypes/subgroups discovery, our framework can help to identify the biomarkers [233, 242, 14] of identified subtypes that can assign a new sample to the correct subgroup/subtype [207].

The distinct subtypes of the same disease cause diverse drug responses [86]. As a result, the treatment choices and the ultimate success for a disease highly depend on its subtype [86, 125, 243]. As future work, one could perform our drug repurposing framework by utilizing the subtypes identified by our clustering framework.

REFERENCES

[1] A. Abdollahi, M. Li, G. Ping, C. Plathow, S. Domhan, F. Kiessling, L. B. Lee, G. McMahon, H.-J. Gröne, K. E. Lipson, et al. Inhibition of platelet-derived growth factor signaling attenuates pulmonary fibrosis. *The Journal of Experimental Medicine*, 201(6):925–935, 2005.

[2] M. Abdul and N. Hoosein. Expression and activity of potassium ion channels in human prostate cancer. *Cancer Letters*, 186(1):99–105, 2002.

[3] V. Abhyankar, P. Bland, and G. Fernandes. The role of systems biologic approach in cell signaling and drug development responses–a mini review. *Medical Sciences*, 6(2):43, 2018.

[4] H. I. Adamali and T. M. Maher. Current and novel drug therapies for idiopathic pulmonary fibrosis. *Drug Design, Development and Therapy*, 6:261, 2012.

[5] C. P. Adams and V. V. Brantner. Estimating the cost of new drug development: is it really $802 million? *Health Affairs*, 25(2):420–428, 2006.

[6] F. Ahangari, C. Becker, D. Foster, M. Chioccioli, M. Nelson, K. Beke, C. Meador, X. Wang, K. Corell, H. Roybal, et al. Saracatinib is a potential novel therapeutic for pulmonary fibrosis. In *B20. THERAPEUTICS" 2020" IN LUNG DISEASE*, pages A2784–A2784. American Thoracic Society, 2020.

[7] J. P. Alao, A. V. Stavropoulou, E. W. Lam, R. C. Coombes, and D. M. Vigushin. Histone deacetylase inhibitor, trichostatin a induces ubiquitin-dependent cyclin d1 degradation in mcf-7 breast cancer cells. *Molecular Cancer*, 5(1):8, 2006.

[8] J. T. Allen and M. A. Spiteri. Growth factors in idiopathic pulmonary fibrosis: relative roles. *Respiratory Research*, 3(1):1, 2001.

[9] R. J. Amato, J. Jac, T. Mohammad, and S. Saxena. Pilot study of rapamycin in patients with hormone-refractory prostate cancer. *Clinical Genitourinary Cancer*, 6(2):97–102, 2008.

[10] E.-a. D. Amir, K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan, and D. Pe'er. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, 31(6):545, 2013.

[11] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[12] T. S. Andrews and M. Hemberg. Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, 2017.

[13] P. Angerer, L. Simon, S. Tritschler, F. A. Wolf, D. Fischer, and F. J. Theis. Single cells make big data: new challenges and opportunities in transcriptomics. *Current Opinion in Systems Biology*, 4:85–91, 2017.

[14] S. Ansari, M. Donato, N. Saberian, and S. Draghici. An approach to infer putative disease-specific mechanisms using neighboring gene networks. *Bioinformatics*, 33(13):1987–1994, 2017.

[15] H. Antoniades, M. Bravo, R. Avila, T. Galanopoulos, J. Neville-Golden, M. Maxwell, and M. Selman. Platelet-derived growth factor in idiopathic pulmonary fibrosis. *Journal of Clinical Investigation*, 86(4):1055, 1990.

[16] T. T. Ashburn and K. B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, 2004.

[17] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[18] D. M. Atienza, C. L. Vogel, B. Trock, and S. M. Swain. Phase II study of oral etoposide for patients with advanced breast cancer. *Cancer*, 76(12):2485–2490, 1995.

[19] M. Baron, A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, A. M. Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Systems*, 3(4):346–360, 2016.

[20] J. B. Bartlett, K. Dredge, and A. G. Dalgleish. The evolution of thalidomide and its IMiD derivatives as anticancer agents. *Nature Reviews Cancer*, 4(4):314–322, 2004.

[21] G. Baruzzo, I. Patuzzi, and B. Di Camillo. Sparsim single cell: a count data simulator for scrna-seq data. *Bioinformatics*, 2019.

[22] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37(1):38, 2019.

[23] C. Behrens, H. Y. Lin, J. J. Lee, M. G. Raso, W. K. Hong, I. I. Wistuba, and R. Lotan. Immunohistochemical expression of basic fibroblast growth factor and fibroblast growth factor receptors 1 and 2 in the pathogenesis of lung cancer. *Clinical Cancer Research*, 14(19):6014–6022, 2008.

[24] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, August 2001.

[25] S. Benzina, J. Harquail, S. Jean, A.-P. Beauregard, C. D Colquhoun, M. Carroll, A. Bos, C. A Gray, and G. A Robichaud. Deoxypodophyllotoxin isolated from juniperus communis induces apoptosis in breast cancer cells. *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)*, 15(1):79–88, 2015.

[26] C. Beyer and J. H. Distler. Tyrosine kinase signaling in fibrotic disorders: translation of basic research to human disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1832(7):897–904, 2013.

[27] P. Bhat-Nakshatri, C. P. Goswami, S. Badve, G. W. Sledge Jr, and H. Nakshatri. Identification of FDA-approved drugs targeting breast cancer stem cells along with biomarkers of sensitivity. *Scientific Reports*, 3, 2013.

[28] F. H. Biase, X. Cao, and S. Zhong. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Research*, 24(11):1787–1796, 2014.

[29] N. Bilani, H. Bahmad, and W. Abou-Kheir. Prostate cancer and aspirin use: Synopsis of the proposed molecular mechanisms. *Frontiers in Pharmacology*, 8, 2017.

[30] D. J. Boffa, F. Luan, D. Thomas, H. Yang, V. K. Sharma, M. Lagman, and M. Suthanthiran. Rapamycin inhibits the growth and metastatic progression of non-small cell lung cancer. *Clinical Cancer Research*, 10(1):293–300, 2004.

[31] B. Booth and R. Zemmel. Prospects for productivity. *Nature Reviews Drug Discovery*, 3(5):451–456, 2004.

[32] D. Bradley, D. Rathkopf, R. Dunn, W. M. Stadler, G. Liu, D. C. Smith, R. Pili, J. Zwiebel, H. Scher, and M. Hussain. Vorinostat in advanced prostate cancer patients progressing on prior chemotherapy (national cancer institute trial 6862). *Cancer*, 115(23):5541–5549, 2009.

[33] G. T. Budd, P. C. Adamson, M. Gupta, P. Homayoun, S. K. Sandstrom, R. F. Murphy, D. McLain, L. Tuason, D. Peereboom, R. M. Bukowski, et al. Phase I/II trial of all-trans retinoic acid and tamoxifen in patients with advanced breast cancer. *Clinical Cancer Research*, 4(3):635–642, 1998.

[34] E. C. Butcher, E. L. Berg, and E. J. Kunkel. Systems biology in drug discovery. *Nature Biotechnology*, 22(10):1253–1259, 2004.

[35] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411, 2018.

[36] C. Cainap, S. Qin, W.-T. Huang, I.-J. Chung, H. Pan, Y. Cheng, M. Kudo, Y.-K. Kang, P.-J. Chen, H. C. Toh, et al. Phase III trial of linifanib versus sorafenib in patients with advanced hepatocellular carcinoma (hcc). In *ASCO Annual Meeting Proceedings*, page 249, 2013.

[37] J. N. Campbell, E. Z. Macosko, H. Fenselau, T. H. Pers, A. Lyubetskaya, D. Tenen, M. Goldman, A. M. Verstegen, J. M. Resch, S. A. McCarroll, et al. A molecular census of arcuate hypothalamus and median eminence cell types. *Nature Neuroscience*, 20(3):484, 2017.

[38] B. Cao, Z. Guo, Y. Zhu, and W. Xu. The potential role of PDGF, IGF-1, TGF-$beta$ expression in idiopathic pulmonary fibrosis. *Chinese Medical Journal*, 113(9):776–

782, 2000.

[39] S. L. Carter, M. M. Centenera, W. D. Tilley, L. A. Selth, and L. M. Butler. Iκbα mediates prostate cancer cell death induced by combinatorial targeting of the androgen receptor. *BMC Cancer*, 16(1):141, 2016.

[40] K. Chamie, P. Ghosh, T. Koppie, V. Romero, C. Troppmann, and R. deVere White. The effect of sirolimus on prostate-specific antigen (psa) levels in male renal transplant recipients without prostate cancer. *American Journal of Transplantation*, 8(12):2668–2673, 2008.

[41] J. Chan. *Statistical Methods for High-throughput Data Integration: Methodologies in Disease Research and Drug Discovery*. PhD thesis, Baylor University, 2019.

[42] J. Chan, X. Wang, J. A. Turner, N. E. Baldwin, and J. Gu. Breaking the paradigm: Dr insight empowers signature-free, enhanced drug repurposing. *Bioinformatics*, 35(16):2818–2826, 2019.

[43] U. R. Chandran, C. Ma, R. Dhir, M. Bisceglia, M. Lyons-Weiler, W. Liang, G. Michalopoulos, M. Becich, and F. A. Monzon. Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer*, 7(1):64, 2007.

[44] B. H. Chao, R. Briesewitz, and M. A. Villalona-Calero. RET fusion genes in non–small-cell lung cancer. *Journal of Clinical Oncology*, 30(35):4439–4441, 2012.

[45] A. Chase, C. Bryant, J. Score, and N. C. Cross. Ponatinib as targeted therapy for fgfr1 fusions associated with the 8p11 myeloproliferative syndrome. *Haematologica*, 98(1):103–106, 2013.

[46] N. Chaudhary, G. Roth, F. Hilberg, J. Müller-Quernheim, A. Prasse, G. Zissel, A. Schnapp, and J. Park. Inhibition of PDGF, VEGF and FGF signalling attenuates fibrosis. *European Respiratory Journal*, 29(5):976–985, 2007.

[47] J. Chen, J. Guo, Z. Chen, J. Wang, M. Liu, and X. Pang. Linifanib (ABT-869) potentiates the efficacy of chemotherapeutic agents through the suppression of receptor tyrosine kinase-mediated AKT/mTOR signaling pathways in gastric cancer. *Scientific Reports*, 6, 2016.

[48] J.-Y. Chen, Y.-A. Tang, W.-S. Li, Y.-C. Chiou, J.-M. Shieh, and Y.-C. Wang. A synthetic podophyllotoxin derivative exerts anti-cancer effects by inducing mitotic arrest and pro-apoptotic ER stress in lung cancer preclinical models. *PloS One*, 8(4):e62082, 2013.

[49] L. Chen, C. Chu, J. Lu, X. Kong, T. Huang, and Y.-D. Cai. Gene ontology and kegg pathway enrichment analysis of a drug target-based classification system. *PloS One*, 10(5):e0126492, 2015.

[50] Y.-L. Chiu, D. M. Carlson, R. S. Pradhan, and J. L. Ricker. Exposure-response (safety) analysis to identify linifanib dose for a phase III study in patients with hepatocellular carcinoma. *Clinical Therapeutics*, 35(11):1770–1777, 2013.

[51] J.-H. Cho, R. Gelinas, K. Wang, A. Etheridge, M. G. Piper, K. Batte, D. Dakhlallah, J. Price, D. Bornman, S. Zhang, et al. Systems biology of interstitial lung diseases: integration of mRNA and microRNA expression changes. *BMC Medical Genomics*, 4(1):1, 2011.

[52] K. S. Choe, J. E. Cowan, J. M. Chan, P. R. Carroll, A. V. D'Amico, and S. L. Liauw. Aspirin use and the risk of prostate cancer mortality in men treated with prostatec-

tomy or radiotherapy. *Journal of Clinical Oncology*, 30(28):3540–3544, 2012.

[53] J. Y. Choi, W. G. Hong, J. H. Cho, E. M. Kim, J. Kim, C.-H. Jung, S.-G. Hwang, H.-D. Um, and J. K. Park. Podophyllotoxin acetate triggers anticancer effects against non-small cell lung cancer cells by promoting cell death via cell cycle arrest, ER stress and autophagy. *International Journal of Oncology*, 47(4):1257–1265, 2015.

[54] C. R. Chong and D. J. Sullivan. New uses for old drugs. *Nature*, 448(7154):645–646, 2007.

[55] C. Ciccarese, F. Massari, R. Iacovelli, M. Fiorentino, R. Montironi, V. Di Nunno, F. Giunchi, M. Brunelli, and G. Tortora. Prostate cancer heterogeneity: Discovering novel molecular targets for therapy. *Cancer Treatment Reviews*, 2017.

[56] J. Cicenas, K. Kalyan, A. Sorokinas, E. Stankunas, J. Levy, I. Meskinyte, V. Stankevicius, A. Kaupinis, and M. Valius. Roscovitine in cancer and other diseases. *Annals of Translational Medicine*, 3(10), 2015.

[57] A. S. Clark, K. A. West, P. M. Blumberg, and P. A. Dennis. Altered protein kinase c (PKC) isoforms in non-small cell lung cancer cells. *Cancer Research*, 63(4):780–786, 2003.

[58] F. S. Collins. Mining for therapeutic gold. *Nature Reviews Drug Discovery*, 10(6):397–397, 2011.

[59] E. Conte, E. Gili, M. Fruciano, M. Korfei, E. Fagone, M. Iemmolo, D. L. Furno, R. Giuffrida, N. Crimi, A. Guenther, et al. PI3K p110γ overexpression in idiopathic pulmonary fibrosis lung tissue and fibroblast cells: in vitro effects of its inhibition. *Laboratory Investigation*, 93(5):566–576, 2013.

[60] F. Conte, G. Fiscon, V. Licursi, D. Bizzarri, T. D'Antò, L. Farina, and P. Paci. A paradigm shift in medicine: a comprehensive review of network-based approaches. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, page 194416, 2019.

[61] H. L. Crowell, C. Soneson, P.-L. Germain, D. Calini, L. Collin, C. Raposo, D. Malhotra, and M. Robinson. On the discovery of population-specific state transitions from multi-sample multi-condition single-cell RNA sequencing data. *BioRxiv*, page 713412, 2019.

[62] S. Dakshanamurthy, N. T. Issa, S. Assefnia, A. Seshasayee, O. J. Peters, S. Madhavan, A. Uren, M. L. Brown, and S. W. Byers. Predicting new indications for approved drugs using a proteo-chemometric method. *Journal of Medicinal Chemistry*, 55(15):6832, 2012.

[63] A. Delfarah, S. Parrish, J. A. Junge, J. Yang, F. Seo, S. Li, J. Mac, P. Wang, S. E. Fraser, and N. A. Graham. Inhibition of nucleotide synthesis promotes replicative senescence of human mammary epithelial cells. *Journal of Biological Chemistry*, 294(27):10564–10578, 2019.

[64] G. D. Demetri, A. T. van Oosterom, C. R. Garrett, M. E. Blackstein, M. H. Shah, J. Verweij, G. McArthur, I. R. Judson, M. C. Heinrich, J. A. Morgan, et al. Efficacy and safety of sunitinib in patients with advanced gastrointestinal stromal tumour after failure of imatinib: a randomised controlled trial. *The Lancet*, 368(9544):1329–1338, 2006.

[65] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*,

343(6167):193–196, 2014.

[66] S. Detchokul and A. G. Frauman. Recent developments in prostate cancer biomarker research: therapeutic implications. *British Journal of Clinical Pharmacology*, 71(2):157–174, 2011.

[67] M. Dickson and J. P. Gagnon. The cost of new drug discovery and development. *Discovery Medicine*, 4(22):172–179, 2009.

[68] J. DiMasi, R. Hansen, and H. Grabowski. The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 22(2):151–186, 2003.

[69] I. Dimopoulou, A. Bamias, P. Lyberopoulos, and M. Dimopoulos. Pulmonary toxicity from novel antineoplastic agents. *Annals of Oncology*, 17(3):372–379, 2005.

[70] S. Draghici. *Statistics and Data Analysis for Microarrays using R and Bioconductor*. Chapman and Hall/CRC Press, 2011.

[71] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichiţa, C. Georgescu, and R. Romero. A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545, 2007.

[72] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichiţa, C. Georgescu, and R. Romero. A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545, 2007.

[73] S. Draghici and T. C. Nguyen. PINS: A Perturbation Clustering Approach for Data Integration and Disease Subtyping, Sept. 15 2016. US Patent App. 15/068,048.

[74] J. T. Dudley, T. Deshpande, and A. J. Butte. Exploiting drug–disease relationships for computational drug repositioning. *Briefings in Bioinformatics*, 2011.

[75] A. Duò, M. D. Robinson, and C. Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7, 2018.

[76] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[77] S. Ekman, M. W. Wynes, and F. R. Hirsch. The mTOR pathway in lung cancer and implications for therapy and biomarker analysis. *Journal of Thoracic Oncology*, 7(6):947–953, 2012.

[78] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[79] E. Fasterius, M. Uhlén, and C. A.-K. Szigyarto. Single-cell RNA-seq variant analysis for exploration of genetic heterogeneity in cancer. *Scientific Reports*, 9(1):9524, 2019.

[80] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. Von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(D1):D808–D815, 2013.

[81] J. F. Gainor and A. T. Shaw. Novel targets in non-small cell lung cancer: ROS1 and RET fusions. *The Oncologist*, 18(7):865–875, 2013.

[82] E. Garattini, G. Paroni, and M. Terao. Retinoids and breast cancer: new clues to increase their activity and selectivity. *Breast Cancer Research*, 14(5):1, 2012.

[83] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–15, Feb 2004.

[84] O. Gautschi, J. Milia, T. Filleron, J. Wolf, D. P. Carbone, D. Owen, R. Camidge, V. Narayanan, R. C. Doebele, B. Besse, et al. Targeting RET in patients with RET-rearranged lung cancers: Results from the global, multicenter RET registry. *Journal of Clinical Oncology*, 35(13):1403–1410, 2017.

[85] Y. Geng, L. Kohli, B. J. Klocke, and K. A. Roth. Chloroquine-induced autophagic vacuole accumulation and cell death in glioma cells is p53 independent. *Neuro-oncology*, 12(5):473–481, 2010.

[86] A. Goldhirsch, W. Wood, A. Coates, R. Gelber, B. Thürlimann, H.-J. Senn, et al. Strategies for subtypes–dealing with the diversity of breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2011. *Annals of Oncology*, page mdr304, 2011.

[87] M. Goolam, A. Scialdone, S. J. Graham, I. C. Macaulay, A. Jedrusik, A. Hupalowska, T. Voet, J. C. Marioni, and M. Zernicka-Goetz. Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell*, 165(1):61–74, 2016.

[88] M. Gordaliza, M. d. Castro, J. Miguel del Corral, and A. S. Feliciano. Antitumor properties of podophyllotoxin and related compounds. *Current Pharmaceutical Design*, 6(18):1811–1839, 2000.

[89] T. Goto, H. Matsushima, Y. Kasuya, Y. Hosaka, T. Kitamura, K. Kawabe, A. Hida, Y. Ohta, T. Simizu, and K. Takeda. The effect of papaverine on morphologic differentiation, proliferation and invasive potential of human prostatic cancer lncap cells. *International Journal of Urology*, 6(6):314–319, 1999.

[90] C. Gridelli, P. Maione, and A. Rossi. The potential role of mTOR inhibitors in non-small cell lung cancer. *The Oncologist*, 13(2):139–147, 2008.

[91] F. Grimminger, A. Günther, and C. Vancheri. The role of tyrosine kinases in the pathogenesis of idiopathic pulmonary fibrosis. *European Respiratory Journal*, pages ERJ–01496, 2015.

[92] F. Grimminger, R. T. Schermuly, and H. A. Ghofrani. Targeting non-malignant disorders with tyrosine kinase inhibitors. *Nature Reviews Drug Discovery*, 9(12):956–970, 2010.

[93] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. van Oudenaarden. Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, 2015.

[94] A. Gucalp, J. A. Sparano, J. Caravelli, J. Santamauro, S. Patil, A. Abbruzzi, C. Pellegrino, J. Bromberg, C. Dang, M. Theodoulou, et al. Phase II trial of saracatinib (AZD0530), an oral SRC-inhibitor for the treatment of patients with hormone receptor-negative metastatic breast cancer. *Clinical Breast Cancer*, 11(5):306–311, 2011.

[95] M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, and Y. Xu. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Computational Biology*, 11(11):e1004575, 2015.

[96] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[97] L. A. Habel, W. Zhao, and J. L. Stanford. Daily aspirin use and prostate cancer risk in a large, multiracial cohort in the us. *Cancer Causes and Control*, 13(5):427–434,

2002.

[98] M. Haigentz Jr, J. Nemunaitis, M. Johnson, N. Mohindra, K. Eaton, M. Patel, M. Awad, D. Briere, N. Sudhakar, D. Faltaos, et al. Phase 1/2 study of mocetinostat and durvalumab (medi4736) in patients with advanced solid tumors and non-small cell lung cancer (NSCLC). In *Journal of Thoracic Oncology*, volume 12, pages S1073–S1074. ELSEVIER SCIENCE INC 360 PARK AVE SOUTH, NEW YORK, NY 10010-1710 USA, 2017.

[99] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1):75, 2017.

[100] C. Hennig. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1):258–271, 2007.

[101] C. Hennig. Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis*, 99(6):1154–1176, 2008.

[102] C. Hennig. *fpc: Flexible procedures for clustering*, 2014. R package version 2.1-7.

[103] F. Hilberg, G. J. Roth, M. Krssak, S. Kautschitsch, W. Sommergruber, U. Tontsch-Grunt, P. Garin-Chesa, G. Bader, A. Zoephel, J. Quant, et al. BIBF 1120: triple angiokinase inhibitor with sustained receptor blockade and good antitumor efficacy. *Cancer Research*, 68(12):4774–4782, 2008.

[104] S. Homma, I. Nagaoka, H. Abe, K. Takahashi, K. Seyama, T. Nukiwa, and S. Kira. Localization of platelet-derived growth factor and insulin-like growth factor I in

the fibrotic lung. *American Journal of Respiratory and Critical Care Medicine*, 152(6):2084–2089, 1995.

[105] A. L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 4(11):682–690, 2008.

[106] A. L. Hopkins et al. Network pharmacology. *Nature Biotechnology*, 25(10):1110–1110, 2007.

[107] H.-S. Hsu, C.-C. Liu, J.-H. Lin, T.-W. Hsu, J.-W. Hsu, K. Su, and S.-C. Hung. Involvement of ER stress, PI3K/AKT activation, and lung fibroblast proliferation in bleomycin-induced pulmonary fibrosis. *Scientific Reports*, 7(1):14272, 2017.

[108] M. Hu, P. Che, X. Han, G.-Q. Cai, G. Liu, V. Antony, T. Luckhardt, G. P. Siegal, Y. Zhou, R.-m. Liu, et al. Therapeutic targeting of src kinase in myofibroblast differentiation and pulmonary fibrosis. *Journal of Pharmacology and Experimental Therapeutics*, 351(1):87–95, 2014.

[109] S. Hu, Q. Zhou, W.-R. Wu, Y.-X. Duan, Z.-Y. Gao, Y.-W. Li, and Q. Lu. Anticancer effect of deoxypodophyllotoxin induces apoptosis of human prostate cancer cells. *Oncology Letters*, 12(4):2918–2923, 2016.

[110] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.

[111] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44, 2009.

[112] H. Huang, L.-J. Li, H.-B. Zhang, and A.-Y. Wei. Papaverine selectively inhibits human prostate cancer cell (PC-3) growth by inducing mitochondrial mediated apoptosis, cell cycle arrest and downregulation of nf-κb/pi3k/akt signalling pathway. *Journal of BU ON.: Official Journal of the Balkan Union of Oncology*, 22(1):112, 2017.

[113] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[114] A. Imrali, X. Mao, M. Yeste-Velasco, J. Shamash, and Y. Lu. Rapamycin inhibits prostate cancer cell growth through cyclin D1 and enhances the cytotoxic efficacy of cisplatin. *American Journal of Cancer Research*, 6(8):1772, 2016.

[115] I. Inoshima, K. Kuwano, N. Hamada, M. Yoshimi, T. Maeyama, N. Hagimoto, Y. Nakanishi, and N. Hara. Induction of CDK inhibitor p21 gene as a new therapeutic strategy against pulmonary fibrosis. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 286(4):L727–L733, 2004.

[116] N. T. Issa, J. Kruger, H. Wathieu, R. Raja, S. W. Byers, and S. Dakshanamurthy. Druggenex-net: a novel computational platform for systems pharmacology and gene expression-based drug repurposing. *BMC Bioinformatics*, 17(1):202, 2016.

[117] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

[118] E. J. Jacobs, C. Rodriguez, A. M. Mondul, C. J. Connell, S. J. Henley, E. E. Calle, and M. J. Thun. A large cohort study of aspirin and other nonsteroidal anti-inflammatory drugs and prostate cancer incidence. *Journal of the National Cancer Institute*, 97(13):975–980, 2005.

[119] E. R. Jang, S.-J. Lim, E. S. Lee, G. Jeong, T.-Y. Kim, Y.-J. Bang, and J.-S. Lee. The histone deacetylase inhibitor trichostatin a sensitizes estrogen receptor α-negative breast cancer cells to tamoxifen. *Oncogene*, 23(9):1724–1736, 2004.

[120] X. Jin, Y. Fang, Y. Hu, J. Chen, W. Liu, G. Chen, M. Gong, P. Wu, T. Zhu, S. Wang, et al. Synergistic activity of the histone deacetylase inhibitor trichostatin a and the proteasome inhibitor ps-341 against taxane-resistant ovarian cancer cell lines. *Oncology Letters*, 13(6):4619–4626, 2017.

[121] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[122] J. H. Joly, A. Delfarah, P. S. Phung, S. Parrish, and N. A. Graham. A synthetic lethal drug combination mimics glucose deprivation–induced cancer cell death in the presence of glucose. *Journal of Biological Chemistry*, 295(5):1350–1365, 2020.

[123] P. Jones, K. Christodoulos, N. Dobbs, P. Thavasu, F. Balkwill, A. Blann, G. Caine, S. Kumar, A. Kakkar, N. Gompertz, et al. Combination antiangiogenesis therapy with marimastat, captopril and fragmin in patients with advanced cancer. *British Journal of Cancer*, 91(1):30–36, 2004.

[124] S. Joost, A. Zeisel, T. Jacob, X. Sun, G. La Manno, P. Lönnerberg, S. Linnarsson, and M. Kasper. Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Cell Systems*, 3(3):221–237, 2016.

[125] S. Jubair, A. Alkhateeb, A. Abou Tabl, L. Rueda, and A. Ngom. A novel approach to identify subtype-specific network biomarkers of breast cancer survivability. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 9(1):1–12, 2020.

[126] T. Kalisky and S. R. Quake. Single-cell genomics. *Nature Methods*, 8(4):311, 2011.

[127] S. Karmakar, Y. Jin, and A. K. Nagaich. Interaction of glucocorticoid receptor (GR) with estrogen receptor (er) α and activator protein 1 (AP1) in dexamethasone-mediated interference of ERα activity. *Journal of Biological Chemistry*, 288(33):24020–24034, 2013.

[128] W. Kassouf, S. Tanguay, and A. G. Aprikian. Nilutamide as second line hormone therapy for prostate cancer after androgen ablation fails. *The Journal of Urology*, 169(5):1742–1744, 2003.

[129] R. E. Kast and M.-E. Halatsch. Matrix metalloproteinase-2 and-9 in glioblastoma: A trio of old drugs–captopril, disulfiram and nelfinavir–are inhibitors with potential as adjunctive treatments in glioblastoma. *Archives of Medical Research*, 43(3):243–247, 2012.

[130] D. Kaushik, V. Vashistha, S. Isharwal, S. A. Sediqe, and M.-F. Lin. Histone deacetylase inhibitors in castration-resistant prostate cancer: molecular mechanism of action and recent clinical trials. *Therapeutic Advances in Urology*, 7(6):388–395, 2015.

[131] G. M. Keating. Nintedanib: a review of its use in patients with idiopathic pulmonary fibrosis. *Drugs*, 75(10):1131–1140, 2015.

[132] B. D. Keith. Systematic review of the clinical effect of glucocorticoids on nonhematologic malignancy. *BMC Cancer*, 8(1):84, 2008.

[133] P. Khatri, M. Sirota, and A. J. Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Computational Biology*, 8(2):e1002375, 2012.

[134] H.-J. Kim and S.-C. Bae. Histone deacetylase inhibitors: molecular mechanisms of action and clinical trials as anti-cancer drugs. *American Journal of Translational*

Research, 3(2):166, 2011.

[135] J. H. Kim and A. R. Scialli. Thalidomide: the tragedy of birth defects and the effective treatment of disease. *Toxicological Sciences*, 122(1):1–6, 2011.

[136] K.-T. Kim, H. W. Lee, H.-O. Lee, H. J. Song, S. Shin, H. Kim, Y. Shin, D.-H. Nam, B. C. Jeong, D. G. Kirsch, et al. Application of single-cell rna sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. *Genome biology*, 17(1):80, 2016.

[137] K.-Y. Kim, H.-J. Cho, S.-N. Yu, S.-H. Kim, H.-S. Yu, Y.-M. Park, N. Mirkheshti, S. Y. Kim, C. S. Song, B. Chatterjee, et al. Interplay of reactive oxygen species, intracellular ca2+ and mitochondrial homeostasis in the apoptosis of prostate cancer cells by deoxypodophyllotoxin. *Journal of Cellular Biochemistry*, 114(5):1124–1134, 2013.

[138] V. Y. Kiselev, T. S. Andrews, and M. Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, page 1, 2019.

[139] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, et al. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5):483, 2017.

[140] V. Y. Kiselev, A. Yiu, and M. Hemberg. scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods*, 15(5):359, 2018.

[141] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[142] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[143] D. Knoerzer, T. Baginski, K. Wade, C. Fan, S. Rapp, K. Regina, F. Shih, M. Burney, S. Rouw, and D. Welsch. Therapeutic efficacy of Sunitinib and other broad spectrum receptor tyrosine kinase inhibitors (RTKI) in bleomycin-induced pulmonary fibrosis. *Journal of Inflammation*, 10(1):1, 2013.

[144] B. Krusche, J. Arend, and T. Efferth. Synergistic inhibition of angiogenesis by artesunate and captopril in vitro and in vivo. *Evidence-Based Complementary and Alternative Medicine*, 2013, 2013.

[145] M.-C. Kuo, S.-J. Chang, and M.-C. Hsieh. Colchicine significantly reduces incident cancer in gout male patients: a 12-year cohort study. *Medicine*, 94(50), 2015.

[146] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wie, S. A. Armstrong, S. Haggarty, P. Clemons, R. Wie, S. Carr, E. Lander, and T. Golub. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, 2006.

[147] A. A. Lane and B. A. Chabner. Histone deacetylase inhibitors in cancer therapy. *Journal of Clinical Oncology*, 27(32):5459–5468, 2009.

[148] P. N. Lara Jr, J. Longmate, C. P. Evans, D. I. Quinn, P. Twardowski, G. Chatta, E. Posadas, W. Stadler, and D. R. Gandara. A phase II trial of the Src-kinase inhibitor AZD0530 in patients with advanced castration-resistant prostate cancer: a california cancer consortium study. *Anti-cancer Drugs*, 20(3):179, 2009.

[149] D. A. Lawson, N. R. Bhakta, K. Kessenbrock, K. D. Prummel, Y. Yu, K. Takai, A. Zhou, H. Eyob, S. Balakrishnan, C.-Y. Wang, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*, 526(7571):131, 2015.

[150] A. Leitch, C. Haslett, and A. Rossi. Cyclin-dependent kinase inhibitor drugs as potential novel anti-inflammatory and pro-resolution agents. *British Journal of Pharmacology*, 158(4):1004–1016, 2009.

[151] H. Li, E. T. Courtois, D. Sengupta, Y. Tan, K. H. Chen, J. J. L. Goh, S. L. Kong, C. Chua, L. K. Hon, W. S. Tan, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics*, 49(5):708, 2017.

[152] H. Li, C. Zhao, Y. Tian, J. Lu, G. Zhang, S. Liang, D. Chen, X. Liu, W. Kuang, and M. Zhu. Src family kinases and pulmonary fibrosis: a review. *Biomedicine & Pharmacotherapy*, page 110183, 2020.

[153] Q. Li, T. Cheng, Y. Wang, and S. H. Bryant. Pubchem as a public resource for drug discovery. *Drug Discovery Today*, 15(23):1052–1057, 2010.

[154] J.-T. Lin, W.-H. Lee, P.-H. Lin, S. W. Haga, Y.-R. Chen, and A. Kranti. A new electron bridge channel 1T-DRAM employing underlap region charge storage. *IEEE Journal of the Electron Devices Society*, 5(1):59–63, 2017.

[155] K.-T. Lin and L.-H. Wang. New dimension of glucocorticoids in cancer treatment. *Steroids*, 111:84–88, 2016.

[156] Z.-Y. Lin, C.-H. Kuo, D.-C. Wu, and W.-L. Chuang. Anticancer effects of clinically acceptable colchicine concentrations on human gastric cancer cell lines. *The Kaohsiung Journal of Medical Sciences*, 32(2):68–73, 2016.

[157] M. Lu, R. Mira-y Lopez, S. Nakajo, K. Nakaya, and Y. Jing. Expression of estrogen receptor α, retinoic acid receptor α and cellular retinoic acid binding protein II genes

is coordinately regulated in human breast cancer cells. *Oncogene*, 24(27):4362–4369, 2005.

[158] A. Lun, D. Risso, and K. Korthauer. SingleCellExperiment: S4 classes for single cell data. *R package version*, 1(0), 2018.

[159] D. Ma, B. Lu, C. Feng, C. Wang, Y. Wang, T. Luo, J. Feng, H. Jia, G. Chi, Y. Luo, and P. Ge. Deoxypodophyllotoxin triggers parthanatos in glioma cells via induction of excessive ROS. *Cancer Letters*, 371(2):194–204, 2016.

[160] Y. Ma and N. Brusselaers. Maintenance use of aspirin or other non-steroidal anti-inflammatory drugs (nsaids) and prostate cancer risk. *Prostate Cancer and Prostatic Diseases*, page 1, 2017.

[161] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[162] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[163] S. K. Madala, S. Schmidt, C. Davidson, M. Ikegami, S. Wert, and W. D. Hardie. MEK-ERK pathway modulation ameliorates pulmonary fibrosis associated with epidermal growth factor receptor activation. *American Journal of Respiratory Cell and Molecular Biology*, 46(3):380–388, 2012.

[164] T. Madej, K. J. Addess, J. H. Fong, L. Y. Geer, R. C. Geer, C. J. Lanczycki, C. Liu, S. Lu, A. Marchler-Bauer, A. R. Panchenko, et al. Mmdb: 3d structures and macromolecular interactions. *Nucleic Acids Research*, 40(D1):D461–D464, 2012.

[165] V. Maire, C. Baldeyron, M. Richardson, B. Tesson, A. Vincent-Salomon, E. Gravier, B. Marty-Prouvost, L. De Koning, G. Rigaill, A. Dumont, et al. TTK/hMPS1 is an attractive therapeutic target for triple-negative breast cancer. *PloS One*, 8(5):e63712, 2013.

[166] V. Maire, F. Némati, M. Richardson, A. Vincent-Salomon, B. Tesson, G. Rigaill, E. Gravier, B. Marty-Prouvost, L. De Koning, G. Lang, et al. Polo-like kinase 1: a potential therapeutic option in combination with conventional chemotherapy for the management of patients with triple-negative breast cancer. *Cancer Research*, 73(2):813–823, 2013.

[167] W. Mangione, Z. Falls, T. Melendy, G. Chopra, and R. Samudrala. Shotgun drug repurposing biotechnology to tackle epidemics and pandemics. *ChemRxiv*, 2020.

[168] H. Mathys, J. Davila-Velderrain, Z. Peng, F. Gao, S. Mohammadi, J. Z. Young, M. Menon, L. He, F. Abdurrob, X. Jiang, et al. Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, page 1, 2019.

[169] C. Mattingly, M. Rosenstein, G. Colby, J. Forrest Jr, and J. Boyer. The comparative toxicogenomics database (CTD): a resource for comparative toxicological studies. *Journal of Experimental Zoology Part A: Comparative Experimental Biology*, 305(9):689–692, 2006.

[170] S. Maubant, B. Tesson, V. Maire, M. Ye, G. Rigaill, D. Gentien, F. Cruzalegui, G. C. Tucker, S. Roman-Roman, and T. Dubois. Transcriptome analysis of Wnt3a-treated triple-negative breast cancer cells. *PloS One*, 10(4):e0122333, 2015.

[171] M. E. Mazzei, L. Richeldi, and H. R. Collard. Nintedanib in the treatment of idiopathic pulmonary fibrosis. *Therapeutic Advances in Respiratory Disease*, 9(3):121–

129, 2015.

[172] D. McCarthy, K. Campbell, A. Lun, and Q. Wills. scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in r. bioRxiv. *DOI: https://doi. org/10.1101/069633*, 2016.

[173] K. McGarry, Y. Graham, S. McDonald, and A. Rashid. Resko: Repositioning drugs by using side effects and knowledge from ontologies. *Knowledge-Based Systems*, 160:34–48, 2018.

[174] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[175] B. H. Mecham, G. T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D. Z. Wetmore, T. J. Mariani, I. S. Kohane, and Z. Szallasi. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Research*, 32(9):74, 2004.

[176] E. B. Meltzer, W. T. Barry, T. A. D'Amico, R. D. Davis, S. S. Lin, M. W. Onaitis, L. D. Morrison, T. A. Sporn, M. P. Steele, and P. W. Noble. Bayesian probit regression model for the diagnosis of pulmonary fibrosis: proof-of-principle. *BMC Medical Genomics*, 4(1):1, 2011.

[177] V. Menon. Clustering single cells: a review of approaches on high-and low-depth single-cell rna-seq data. *Briefings in Functional Genomics*, 17(4):240–245, 2017.

[178] P. F. Mercer, H. V. Woodcock, J. D. Eley, M. Platé, M. G. Sulikowski, P. F. Durrenberger, L. Franklin, C. B. Nanthakumar, Y. Man, F. Genovese, et al. Exploration of

a potent PI3 kinase/mTOR inhibitor as a novel anti-fibrotic agent in IPF. *Thorax*, pages thoraxjnl–2015, 2016.

[179] W. Messersmith, S. Nallapareddy, J. Arcaroli, A. Tan, N. Foster, J. Wright, J. Picus, B. Goh, M. Hidalgo, and C. Erlichman. A phase II trial of saracatinib (AZD0530), an oral src inhibitor, in previously treated metastatic pancreatic cancer. In *ASCO Annual Meeting Proceedings*, volume 28, page e14515, 2010.

[180] C. Mitrea, Z. Taghavi, B. Bokanizad, S. Hanoudi, R. Tagett, M. Donato, C. Voichiţa, and S. Draghici. Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*, 4:278, 2013.

[181] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1-2):91–118, 2003.

[182] M. Mottamal, S. Zheng, T. L. Huang, and G. Wang. Histone deacetylase inhibitors in clinical studies as templates for new anticancer agents. *Molecules*, 20(3):3898–3941, 2015.

[183] R. J. Motzer, T. E. Hutson, P. Tomczak, M. D. Michaelson, R. M. Bukowski, O. Rixe, S. Oudard, S. Negrier, C. Szczylik, S. T. Kim, et al. Sunitinib versus interferon alfa in metastatic renal-cell carcinoma. *New England Journal of Medicine*, 356(2):115–124, 2007.

[184] E. Nakajima, B. Helfrich, D. Chan, Z. Zhang, F. Hirsch, V. Chen, D. Ma, and P. Bunn Jr. Enzastaurin a protein kinase cbeta-selective inhibitor, inhibits the growth of sclc and nsclc cell lines. *Journal of Clinical Oncology*, 24(18_suppl):13138–13138, 2006.

[185] S. Namazi, J. Rostami-Yalmeh, E. Sahebi, M. Jaberipour, M. Razmkhah, and A. Hosseini. The role of captopril and losartan in prevention and regression of tamoxifen-induced resistance of breast cancer cell line MCF-7: an in vitro study. *Biomedicine & Pharmacotherapy*, 68(5):565–571, 2014.

[186] N. E. Navin. The first five years of single-cell cancer genomics and beyond. *Genome Research*, 25(10):1499–1507, 2015.

[187] J. W. Neal and L. V. Sequist. Complex role of histone deacetylase inhibitors in the treatment of non–small-cell lung cancer. *Journal of Clinical Oncology*, 30(18):2280–2282, 2012.

[188] J. Nelson and R. E. Harris. Inverse association of prostate cancer and non-steroidal anti-inflammatory drugs (NSAIDs): results of a case-control study. *Oncology Reports*, 7(1):169–239, 2000.

[189] S. Novello, C. Camps, F. Grossi, J. Mazieres, L. Abrey, J.-M. Vernejoux, A. Thall, S. Patyna, T. Usari, Z. Wang, et al. Phase II study of sunitinib in patients with non-small cell lung cancer and irradiated brain metastases. *Journal of Thoracic Oncology*, 6(7):1260–1266, 2011.

[190] M. Núñez, V. Medina, G. Cricco, M. Croci, C. Cocca, E. Rivera, R. Bergoc, and G. Martín. Glibenclamide inhibits cell growth by inducing g0/g1 arrest in the human breast cancer cell line mda-mb-231. *BMC Pharmacology and Toxicology*, 14(1):6, 2013.

[191] H. B. Nygaard, A. F. Wagner, G. S. Bowen, S. P. Good, M. G. MacAvoy, K. A. Strittmatter, A. C. Kaufman, B. J. Rosenberg, T. Sekine-Konno, P. Varma, et al. A phase Ib multiple ascending dose study of the safety, tolerability, and central nervous system

availability of AZD0530 (saracatinib) in Alzheimer's disease. *Alzheimer's Research & Therapy*, 7(1):1, 2015.

[192] Y. Oh, R. S. Herbst, H. Burris, A. Cleverly, L. Musib, M. Lahn, and G. Bepler. Enzastaurin, an oral serine/threonine kinase inhibitor, as second-or third-line therapy of non–small-cell lung cancer. *Journal of Clinical Oncology*, 26(7):1135–1141, 2008.

[193] K. C. Olsen, A. P. Epa, A. A. Kulkarni, R. M. Kottmann, C. E. McCarthy, G. V. Johnson, T. H. Thatcher, R. P. Phipps, and P. J. Sime. Inhibition of transglutaminase 2, a novel target for pulmonary fibrosis, by two small electrophilic molecules. *American Journal of Respiratory Cell and Molecular Biology*, 50(4):737–747, 2014.

[194] T. K. Olsen and N. Baryawno. Introduction to single-cell RNA sequencing. *Current Protocols in Molecular Biology*, 122(1):e57, 2018.

[195] P.-S. Ong, L. Wang, D. M.-H. Chia, J. Y.-X. Seah, L.-R. Kong, W.-L. Thuya, A. Chinnathambi, J.-Y. A. Lau, A. L.-A. Wong, W.-P. Yong, et al. A novel combinatorial strategy using Seliciclib® and belinostat® for eradication of non-small cell lung cancer via apoptosis induction and BID activation. *Cancer Letters*, 381(1):49–57, 2016.

[196] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins. How many drug targets are there? *Nature Reviews Drug Discovery*, 5(12):993–996, 2006.

[197] C. Pacini. *DrugVsDisease: Comparison of disease and drug profiles using Gene set Enrichment Analysis*. Bioinformatics, 2013. R package version 2.4.0.

[198] Y. Pan, T. Cheng, Y. Wang, and S. H. Bryant. Pathway analysis for drug repositioning based on public database mining. *Journal of Chemical Information and Modeling*, 54(2):407–418, 2014.

[199] G. Paroni, M. Fratelli, G. Gardini, C. Bassano, M. Flora, A. Zanetti, V. Guarnaccia, P. Ubezio, F. Centritto, M. Terao, et al. Synergistic antitumor activity of lapatinib and retinoids on a novel subtype of breast cancer with coamplification of ERBB2 and RARA. *Oncogene*, 31(29):3431–3443, 2012.

[200] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suvà, A. Regev, and B. E. Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, June 2014.

[201] N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006.

[202] L. Payen, L. Delugin, A. Courtois, Y. Trinquart, A. Guillouzo, and O. Fardel. The sulphonylurea glibenclamide inhibits multidrug resistance protein (mrp1) activity in human lung cancer cells. *British Journal of Pharmacology*, 132(3):778–784, 2001.

[203] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[204] R. Peng, S. Sridhar, G. Tyagi, J. E. Phillips, R. Garrido, P. Harris, L. Burns, L. Renteria, J. Woods, L. Chen, et al. Bleomycin induces molecular changes directly relevant to idiopathic pulmonary fibrosis: a model for "active" disease. *PloS One*, 8(4):e59348, 2013.

[205] A. Peyvandipour, N. Saberian, A. Shafi, M. Donato, and S. Draghici. A novel com-

putational approach for drug repurposing using systems biology. *Bioinformatics*, 34(16):2817–2825, August 2018.

[206] A. Peyvandipour, A. Shafi, N. Saberian, and S. Draghici. Identification of cell types from single cell data using stable clustering. *Scientific Reports*, 10(1):1–12, 2020.

[207] H. Q. Pham, L. Rueda, and A. Ngom. A data integration approach for detecting biomarkers of breast cancer survivability. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 49–60. Springer, 2020.

[208] E. Pierson and C. Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16:241, 2015.

[209] A. Planche, M. Bacac, P. Provero, C. Fusco, M. Delorenzi, J.-C. Stehle, and I. Stamenkovic. Identification of prognostic molecular features in the reactive stroma of human breast and prostate cancer. *PloS One*, 6(5):18640, 2011.

[210] A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, N. Ramalingam, G. Sun, M. Thu, M. Norris, R. Lebofsky, D. Toppani, D. W. Kemp Ii, M. Wong, B. Clerkson, B. N. Jones, S. Wu, L. Knutsson, B. Alvarado, J. Wang, L. S. Weaver, A. P. May, R. C. Jones, M. A. Unger, A. R. Kriegstein, and J. A. A. West. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, 32(10):1053–1058, Oct. 2014.

[211] C. J. Poole, A. Lisyanskaya, S. Rodenhuis, G. Kristensen, E. P. Lauraine, M. Cantarini, U. Emeribe, M. Stuart, and I. Coquard. A randomized phase II clinical trial of the src inhibitor saracatinib (AZD0530) and carboplatin plus paclitaxel (C plus P) versus C

plus p in patients (PTS) wiht advanced platinum-sensitive epithelial ovarian cancer (EOC). *Annals of Oncology*, 21(Supplement 8):304–305, 2010.

[212] K. A. Price, C. G. Azzoli, L. M. Krug, M. C. Pietanza, N. A. Rizvi, W. Pao, M. G. Kris, G. J. Riely, R. T. Heelan, M. E. Arcila, et al. Phase II trial of gefitinib and everolimus in advanced non-small cell lung cancer. *Journal of Thoracic Oncology*, 5(10):1623–1629, 2010.

[213] X. Qian, J. Li, J. Ding, Z. Wang, L. Duan, and G. Hu. Glibenclamide exerts an antitumor activity through reactive oxygen species–c-jun nh (2)-terminal kinase pathway in human gastric cancer cell line mgc-803. *Biochemical Pharmacology*, 76(12):1705–1715, 2008.

[214] G. Raghu and M. Selman. Nintedanib and Pirfenidone. new antifibrotic treatments indicated for idiopathic pulmonary fibrosis offer hopes and raises questions. *American Journal of Respiratory and Critical Care Medicine*, 191(3):252–254, 2015.

[215] J. S. Rai, M. J. Henley, and H. L. Ratan. Mammalian target of rapamycin: a new target in prostate cancer. In *Urologic Oncology: Seminars and Original Investigations*, volume 28, pages 134–138. Elsevier, 2010.

[216] H. R. Rajpura and A. Ngom. Drug target interaction predictions using pu-leaming under different experimental setting for four formulations namely known drug target pair prediction, drug prediction, target prediction and unknown drug target pair prediction. In *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7. IEEE, 2018.

[217] A. S. Reddy and S. Zhang. Polypharmacology: drug discovery for the future. *Expert Review of Clinical Pharmacology*, 6(1):41–47, 2013.

[218] M. Ren, M. Hong, G. Liu, H. Wang, V. Patel, P. Biddinger, J. Silva, J. Cowell, and Z. Hao. Novel fgfr inhibitor ponatinib suppresses the growth of non-small cell lung cancer cells overexpressing fgfr1. *Oncology Reports*, 29(6):2181–2190, 2013.

[219] E. A. Renzoni, D. J. Abraham, S. Howat, X. Shi-Wen, P. Sestini, G. Bou-Gharios, A. U. Wells, S. Veeraraghavan, A. G. Nicholson, C. P. Denton, et al. Gene expression profiling reveals novel TGFβ targets in adult lung fibroblasts. *Respiratory Research*, 5(1):24, 2004.

[220] C. K. Rhee, S. H. Lee, H. K. Yoon, S. C. Kim, S. Y. Lee, S. S. Kwon, Y. K. Kim, K. H. Kim, T. J. Kim, and J. W. Kim. Effect of nilotinib on bleomycin-induced acute lung injury and pulmonary fibrosis in mice. *Respiration*, 82(3):273–287, 2011.

[221] L. V. Rhodes, A. M. Nitschke, H. C. Segar, E. C. Martin, J. L. Driver, S. Elliott, S. Y. Nam, M. Li, K. P. Nephew, M. E. Burow, et al. The histone deacetylase inhibitor trichostatin a alters microrna expression profiles in apoptosis-resistant breast cancer cells. *Oncology Reports*, 27(1):10–16, 2012.

[222] L. Richeldi, U. Costabel, M. Selman, D. S. Kim, D. M. Hansell, A. G. Nicholson, K. K. Brown, K. R. Flaherty, P. W. Noble, G. Raghu, M. Brun, A. Gupta, N. Juhel, M. Klüglich, and R. M. Bois. Efficacy of a tyrosine kinase inhibitor in idiopathic pulmonary fibrosis. *New England Journal of Medicine*, 365(12):1079–1087, 2011.

[223] M. Roh, C. Kim, B. Park, G. Kim, J. Jeong, H. Kwon, D. Suh, K. Cho, S.-B. Yee, and Y. Yoo. Mechanism of histone deacetylase inhibitor trichostatin a induced apoptosis in human osteosarcoma cells. *Apoptosis*, 9(5):583–589, 2004.

[224] Z. Rong, L. Li, F. Fei, L. Luo, and Y. Qu. Combined treatment of glibenclamide and cocl2 decreases mmp9 expression and inhibits growth in highly metastatic breast

cancer. *Journal of Experimental & Clinical Cancer Research*, 32(1):32, 2013.

[225] H. K. Rooprai, A. Kandanearatchi, S. Maidment, M. Christidou, G. Trillo-Pazos, D. T. Dexter, G. Rucklidge, W. Widmer, and G. J. Pilkington. Evaluation of the effects of swainsonine, captopril, tangeretin and nobiletin on the biological behaviour of brain tumour cells in vitro. *Neuropathology and Applied Neurobiology*, 27(1):29–39, 2001.

[226] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.

[227] S. I. Rothschild and O. Gautschi. Src tyrosine kinase inhibitors in the treatment of lung cancer: rationale and clinical data. *Clinical Investigation*, 2(4):387–396, 2012.

[228] S. I. Rothschild, O. Gautschi, E. B. Haura, and F. M. Johnson. Src inhibitors in lung cancer: current status and future directions. *Clinical Lung Cancer*, 11(4):238–242, 2010.

[229] P. M. Rothwell, F. G. R. Fowkes, J. F. Belch, H. Ogawa, C. P. Warlow, and T. W. Meade. Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. *The Lancet*, 377(9759):31–41, 2011.

[230] D. Rundle-Thiele, R. Head, L. Cosgrove, and J. H. Martin. Repurposing some older drugs that cross the blood–brain barrier and have potential anticancer activity to provide new treatment options for glioblastoma. *British Journal of Clinical Pharmacology*, 81(2):199–209, 2016.

[231] A. Saadatpour, S. Lai, G. Guo, and G.-C. Yuan. Single-cell analysis in cancer genomics. *Trends in Genetics*, 31(10):576–586, 2015.

[232] N. Saberian, A. Peyvandipour, M. Donato, S. Ansari, and S. Draghici. A new computational drug repurposing method using established disease–drug pair knowledge. *Bioinformatics*, 35(19):3672–3678, 2019.

[233] N. Saberian, A. Shafi, A. Peyvandipour, and S. Draghici. MAGPEL: an automated pipeline for inferring variant-driven gene panels from the full-length biomedical literature. *Scientific Reports*, 10(1):1–11, 2020.

[234] J. N. Sarkaria, P. Schwingler, S. E. Schild, P. T. Grogan, A. C. Mladek, S. J. Mandrekar, A. D. Tan, T. Kobayashi, R. S. Marks, H. Kita, et al. Phase i trial of sirolimus combined with radiation and cisplatin in non-small cell lung cancer. *Journal of Thoracic Oncology*, 2(8):751–757, 2007.

[235] H. Sasaki, M. Shitara, K. Yokota, Y. Hikosaka, S. Moriyama, M. Yano, and Y. Fujii. Increased fgfr1 copy number in lung squamous cell carcinomas. *Molecular Medicine Reports*, 5(3):725–728, 2012.

[236] R. Savai, S. S. Pullamsetti, G.-A. Banat, N. Weissmann, H. A. Ghofrani, F. Grimminger, and R. T. Schermuly. Targeting cancer with phosphodiesterase inhibitors. *Expert Opinion on Investigational Drugs*, 19(1):117–131, 2010.

[237] A. Saxena. Drug targets for COVID-19 therapeutics: Ongoing global efforts. *Journal of Biosciences*, 45(1):1–24, 2020.

[238] H. W. Schroeder. Mixing the old with the new: Drug repurposing for immune deficiency in the era of precision medicine and pediatric genomics. *The Journal of Allergy and Clinical Immunology: In Practice*, 6(6):2168–2169, 2018.

[239] J. Schuler and R. Samudrala. Fingerprinting cando: Increased accuracy with structure-and ligand-based shotgun drug repurposing. *ACS omega*, 4(17):17393–17403, 2019.

[240] Å. Segerstolpe, A. Palasantza, P. Eliasson, E.-M. Andersson, A.-C. Andréasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metabolism*, 24(4):593–607, 2016.

[241] S. A. Selamat, B. S. Chung, L. Girard, W. Zhang, Y. Zhang, M. Campan, K. D. Siegmund, M. N. Koss, J. A. Hagen, W. L. Lam, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Research*, 22(7):1197–1211, 2012.

[242] A. Shafi, T. Nguyen, A. Peyvandipour, and S. Draghici. GSMA: an approach to identify robust global and test gene signatures using meta-analysis. *Bioinformatics*, 1:1–9, 2019.

[243] A. Shafi, T. Nguyen, A. Peyvandipour, H. Nguyen, and S. Draghici. A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures. *Frontiers in Genetics*, 10:159, 2019.

[244] A. K. Shalek and M. Benson. Single-cell analyses to tailor treatments. *Science Translational Medicine*, 9(408), 2017.

[245] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[246] A. Sharma, A. Jacob, M. Tandon, and D. Kumar. Orphan drug: Development trends and strategies. *Journal of Pharmacy and Bioallied Sciences*, 2(4):290, 2010.

[247] N. Sharma, S. Thomas, E. B. Golden, F. M. Hofman, T. C. Chen, N. A. Petasis, A. H. Schönthal, and S. G. Louie. Inhibition of autophagy and induction of breast cancer cell death by mefloquine, an antimalarial agent. *Cancer Letters*, 326(2):143–154, 2012.

[248] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science Translational Medicine*, 3(96), 2011.

[249] C. Skriver, C. Dehlendorff, M. Borre, K. Brasso, H. T. Sørensen, J. Hallas, S. B. Larsen, A. Tjønneland, and S. Friis. Low-dose aspirin or other nonsteroidal anti-inflammatory drug use and prostate cancer risk: a nationwide study. *Cancer Causes & Control*, 27(9):1067–1079, 2016.

[250] G. Smyth. *Limma: Linear models for microarray data. In: Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer: New York, 2005.

[251] M. Socinski, S. Novello, J. Sanchez, J. Brahmer, R. Govindan, C. Belani, J. Atkins, H. Gillenwater, C. Palleres, and R. Chao. Efficacy and safety of sunitinib in previously treated, advanced non-small cell lung cancer (nsclc): Preliminary results of a multicenter phase ii trial. *Journal of Clinical Oncology*, 24(18_suppl):7001–7001, 2006.

[252] J.-C. Soria, F. Shepherd, J.-Y. Douillard, J. Wolf, G. Giaccone, L. Crino, F. Cappuzzo, S. Sharma, S. Gross, S. Dimitrijevic, et al. Efficacy of everolimus (RAD001) in patients with advanced NSCLC previously treated with chemotherapy alone or with

chemotherapy and EGFR inhibitors. *Annals of Oncology*, 20(10):1674–1681, 2009.

[253] H. Soule, J. Vazquez, A. Long, S. Albert, and M. Brennan. A human cell line from a pleural effusion derived from a breast carcinoma. *Journal of the National Cancer Institute*, 51(5):1409–1416, 1973.

[254] P. Stopfer, K. Rathgen, D. Bischoff, S. Lüdtke, K. Marzin, R. Kaiser, K. Wagner, and T. Ebner. Pharmacokinetics and metabolism of BIBF 1120 after oral dosing to healthy male volunteers. *Xenobiotica*, 41(4):297–311, 2011.

[255] Y. Sun, X. Lin, and H. Chang. Proliferation inhibition and apoptosis of breast cancer mcf-7 cells under the influence of colchicine. *J. BUON*, 3:570–575, 2016.

[256] L. M. Sutton, M. A. Warmuth, W. P. Petros, and E. P. Winer. Pharmacokinetics and clinical impact of all-trans retinoic acid in metastatic breast cancer: a phase ii trial. *Cancer Chemotherapy and Pharmacology*, 40(4):335–341, 1997.

[257] N. Takai and H. Narahara. Preclinical studies of chemotherapy using histone deacetylase inhibitors in endometrial cancer. *Obstetrics and Gynecology International*, 2010, 2010.

[258] T. Takeuchi, S. Tomida, Y. Yatabe, T. Kosaka, H. Osada, K. Yanagisawa, T. Mitsudomi, and T. Takahashi. Expression profile–defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. *Journal of Clinical Oncology*, 24(11):1679–1688, 2006.

[259] S. Tavakoli, A. Hajibagheri, and G. Sukthankar. Learning social graph topologies using generative adversarial neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction*, 2017.

[260] S. Tavakoli and S. Yooseph. Learning a mixture of microbial networks using minorization–maximization. *Bioinformatics*, 35(14):i23–i30, 2019.

[261] C. Tekle, E. Giovannetti, J. Sigmond, J. Graff, K. Smid, and G. Peters. Molecular pathways involved in the synergistic interaction of the pkcβ inhibitor enzastaurin with the antifolate pemetrexed in non-small cell lung cancer cells. *British Journal of Cancer*, 99(5):750–759, 2008.

[262] A. Thakur, M. Goldbaum, and S. Yousefi. Convex representations using deep archetypal analysis for predicting glaucoma. *IEEE Journal of Translational Engineering in Health and Medicine*, 8:1–7, 2020.

[263] A. Thomas, S. V. Liu, D. S. Subramaniam, and G. Giaccone. Refining the treatment of NSCLC according to histological and molecular subtypes. *Nature Reviews Clinical Oncology*, 12(9):511–526, 2015.

[264] C. A. Tracy and H. Widom. Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics*, 159(1):151–174, 1994.

[265] B. Tran, D. Tran, H. Nguyen, N. S. Vo, and T. Nguyen. Ria: a novel regression-based imputation approach for single-cell RNA sequencing. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–9. IEEE, 2019.

[266] C. Trapnell. Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498, 2015.

[267] B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371, 2014.

[268] G. C. Tseng and W. H. Wong. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, 61(1):10–16, 2005.

[269] P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*, 7:39921, 2017.

[270] B. Turanli, O. Altay, J. Borén, H. Turkez, J. Nielsen, M. Uhlen, K. Y. Arga, and A. Mardinoglu. Systems biology based drug repositioning for development of cancer therapy. In *Seminars in Cancer Biology*. Elsevier, 2019.

[271] J. Vansteenkiste, B. Solomon, M. Boyer, J. Wolf, N. Miller, L. Di Scala, I. Pylvaenaeinen, K. Petrovic, S. Dimitrijevic, B. Anrys, et al. Everolimus in combination with pemetrexed in patients with advanced non-small cell lung cancer previously treated with chemotherapy: a phase i study using a novel, adaptive bayesian dose-escalation model. *Journal of Thoracic Oncology*, 6(12):2120–2129, 2011.

[272] D. M. Vigushin, S. Ali, P. E. Pace, N. Mirsaidi, K. Ito, I. Adcock, and R. C. Coombes. Trichostatin a is a histone deacetylase inhibitor with potent antitumor activity against breast cancer in vivo. *Clinical Cancer Research*, 7(4):971–976, 2001.

[273] M. A. Villarreal. Orphan drug act: background and proposed legislation in the 107th congress. Congressional Research Service, the Library of Congress, 2001.

[274] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080, 2009.

[275] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal*

of Machine Learning Research, 11(Oct):2837–2854, 2010.

[276] C. Voichiţa and S. Draghici. *ROntoTools: R Onto-Tools suite*, 2013. R package.

[277] T. A. Wallace, R. L. Prueitt, M. Yi, T. M. Howe, J. W. Gillespie, H. G. Yfantis, R. M. Stephens, N. E. Caporaso, C. A. Loffredo, and S. Ambs. Tumor immunobiological differences in prostate cancer between African-American and European-American men. *Cancer Research*, 68(3):927–936, 2008.

[278] C. A. Walsh, J. C. Bolger, C. Byrne, S. Cocchiglia, Y. Hao, A. Fagan, L. Qin, A. Cahalin, D. McCartan, M. McIlroy, et al. Global gene repression by the steroid receptor coactivator SRC-1 promotes oncogenesis. *Cancer Research*, 74(9):2533–2544, 2014.

[279] E. S. Wang, K. Yee, L. P. Koh, D. Hogge, S. Enschede, D. M. Carlson, M. Dudley, K. Glaser, E. McKeegan, D. H. Albert, et al. Phase 1 trial of linifanib (ABT-869) in patients with refractory or relapsed acute myeloid leukemia. *Leukemia & Lymphoma*, 53(8):1543–1551, 2012.

[280] Y. Wang and N. E. Navin. Advances and applications of single-cell sequencing technologies. *Molecular Cell*, 58(4):598–609, 2015.

[281] Y. Wang, J. K. Yella, S. Ghandikota, T. C. Cherukuri, S. K. Madala, and A. G. Jegga. Pan-transcriptome-based candidate therapeutic discovery for idiopathic pulmonary fibrosis. *bioRxiv*, page 824367, 2019.

[282] S. N. Waqar, P. K. Gopalan, K. Williams, S. Devarakonda, and R. Govindan. A phase I trial of sunitinib and rapamycin in patients with advanced non-small cell lung cancer. *Chemotherapy*, 59(1):8–13, 2013.

[283] J. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[284] C.-C. Wen, H.-M. Chen, S.-S. Chen, L.-T. Huang, W.-T. Chang, W.-C. Wei, L.-C. Chou, P. Arulselvan, J.-B. Wu, S.-C. Kuo, et al. Specific microtubule-depolymerizing agents augment efficacy of dendritic cell-based cancer vaccines. *Journal of Biomedical Science*, 18(1):44, 2011.

[285] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83, 1945.

[286] M. D. Wilkerson and D. N. Hayes. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12):1572–1573, 2010.

[287] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(suppl 1):D668–D672, 2006.

[288] L. Wollin, I. Maillet, V. Quesniaux, A. Holweg, and B. Ryffel. Antifibrotic and anti-inflammatory activity of the tyrosine kinase inhibitor nintedanib in experimental models of lung fibrosis. *Journal of Pharmacology and Experimental Therapeutics*, 349(2):209–220, 2014.

[289] L. Wollin, E. Wex, A. Pautsch, G. Schnapp, K. E. Hostettler, S. Stowasser, and M. Kolb. Mode of action of nintedanib in the treatment of idiopathic pulmonary fibrosis. *European Respiratory Journal*, pages ERJ–01749, 2015.

[290] J. E. Wooldridge, C. M. Anderson, and M. C. Perry. Corticosteroids in advanced cancer. *Oncology (Williston Park, NY)*, 15(2):225–34, 2001.

[291] C. Xu and Z. Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, page btv088, 2015.

[292] P. Xuan, Y. Cao, T. Zhang, X. Wang, S. Pan, and T. Shen. Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics*, 35(20):4108–4119, 2019.

[293] G. Yan, K. Graham, and S. Lanza-Jacoby. Curcumin enhances the anticancer effects of trichostatin a in breast cancer cells. *Molecular Carcinogenesis*, 52(5):404–411, 2013.

[294] K.-H. Yan, Y.-W. Lin, C.-H. Hsiao, Y.-C. Wen, K.-H. Lin, C.-C. Liu, M.-C. Hsieh, C.-J. Yao, M.-D. Yan, G.-M. Lai, S.-E. CHUANG, and L.-M. LEE. Mefloquine induces cell death in prostate cancer cells and provides a potential novel treatment strategy in vivo. *Oncology Letters*, 5(5):1567–1571, 2013.

[295] K.-H. Yan, C.-J. Yao, C.-H. Hsiao, K.-H. Lin, Y.-W. Lin, Y.-C. Wen, C.-C. Liu, M.-D. Yan, S.-E. Chuang, G.-M. Lai, et al. Mefloquine exerts anticancer activity in prostate cancer cells via ROS-mediated modulation of Akt, ERK, jnk and AMPK signaling. *Oncology Letters*, 5(5):1541–1545, 2013.

[296] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, et al. Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural and Molecular Biology*, 20(9):1131, 2013.

[297] G. Yang, A. Ma, and Z. S. Qin. An integrated system biology approach yields drug repositioning candidates for the treatment of heart failure. *Frontiers in Genetics*, 10:916, 2019.

[298] X. Yang, A. T. Ferguson, S. J. Nass, D. L. Phillips, K. A. Butash, S. M. Wang, J. G. Herman, and N. E. Davidson. Transcriptional activation of estrogen receptor α

in human breast cancer cells by histone deacetylase inhibition. *Cancer Research*, 60(24):6890–6894, 2000.

[299] T. Yasukagawa, Y. Niwa, S. Simizu, and K. Umezawa. Suppression of cellular invasion by glybenclamide through inhibited secretion of platelet-derived growth factor in ovarian clear cell carcinoma ES-2 cells. *FEBS Letters*, 586(10):1504–1509, 2012.

[300] M. A. Yıldırım, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal. Drug-target network. *Nature Biotechnology*, 25(10):1119–1126, 2007.

[301] K. Yoshida, K. Kuwano, N. Hagimoto, K. Watanabe, T. Matsuba, M. Fujita, I. Inoshima, and N. Hara. MAP kinase activation and apoptosis in lung tissues from patients with idiopathic pulmonary fibrosis. *The Journal of Pathology*, 198(3):388–396, 2002.

[302] S. Yousefi, T. Elze, L. Pasquale, O. Saeedi, M. Wang, L. Shen, S. Wellik, C. G. De Moraes, J. S. Myers, and M. V. Boland. Clinical utility of the artificial intelligence enabled dashboard for glaucoma monitoring. *Investigative Ophthalmology & Visual Science*, 61(7):4526–4526, 2020.

[303] Y. P. Yu, D. Landsittel, L. Jing, J. Nelson, B. Ren, L. Liu, C. McDonald, R. Thomas, R. Dhir, S. Finkelstein, et al. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *Journal of Clinical Oncology*, 22(14):2790–2799, 2004.

[304] G.-C. Yuan, L. Cai, M. Elowitz, T. Enver, G. Fan, G. Guo, R. Irizarry, P. Kharchenko, J. Kim, S. Orkin, et al. Challenges and emerging directions in single-cell analysis. *Genome Biology*, 18(1):84, 2017.

[305] P. Yuan, L. Di, X. Zhang, M. Yan, D. Wan, L. Li, Y. Zhang, J. Cai, H. Dai, Q. Zhu, et al. Efficacy of oral etoposide in pretreated metastatic breast cancer: A multicenter phase 2 study. *Medicine*, 94(17), 2015.

[306] P. Yuan, B.-H. Xu, J.-Y. Wang, F. Ma, Y. Fan, Q. Li, and P. Zhang. Oral etoposide monotherapy is effective for metastatic breast cancer with heavy prior therapy. *Chinese Medical Journal*, 125(5):775–779, 2012.

[307] L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174, 2017.

[308] L. Zappia, B. Phipson, and A. Oshlack. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Computational Biology*, 14(6):e1006245, 2018.

[309] J. Zhang, S. Kalyankrishna, M. Wislez, N. Thilaganathan, B. Saigal, W. Wei, L. Ma, I. I. Wistuba, F. M. Johnson, and J. M. Kurie. SRC-family kinases are activated in non-small cell lung cancer and promote the survival of epidermal growth factor receptor-dependent cell lines. *The American Journal of Pathology*, 170(1):366–376, 2007.

[310] Y. Zhou, H. Peng, H. Sun, X. Peng, C. Tang, Y. Gan, X. Chen, A. Mathur, B. Hu, M. D. Slade, et al. Chitinase 3–like 1 suppresses injury and promotes fibroproliferative responses in mammalian lung fibrosis. *Science Translational Medicine*, 6(240):240ra76–240ra76, 2014.

ABSTRACT

TOWARDS PERSONALIZED MEDICINE: COMPUTATIONAL APPROACHES FOR DRUG REPURPOSING AND CELL TYPE IDENTIFICATION

by

AZAM PEYVANDIPOUR

December 2020

Advisor: Dr. Sorin Draghici

Major: Computer Science

Degree: Doctor of Philosophy

The traditional drug discovery process is extremely slow and costly. More than 90% of drugs fail to pass beyond the early stage of development and toxicity tests, and many of the drugs that go through early phases of the clinical trials fail because of adverse reactions, side effects, or lack of efficiency. In spite of unprecedented investments in research and development (R&D), the number of new FDA-approved drugs remains low, reflecting the limitations of the current R&D model. In this context, finding new disease indications for existing drugs sidesteps these issues and can therefore increase the available therapeutic choices at a fraction of the cost of new drug development. In this thesis, we introduce a drug repurposing approach that takes advantage of prior knowledge of drug targets, disease-related genes, and signaling pathways to construct a drug-disease network composed of the genes that are most likely perturbed by a drug. Systems biology can be used as an effective platform in drug discovery and development by leveraging the understanding of interactions between the different system components [34, 141, 216]. By performing a system-level analysis on this network, our approach estimates the amount of perturbation caused by drugs and diseases and discovers drugs with the potential desired effects on

the given disease. Next, we develop a stable clustering method that employs a bootstrap approach to identify the stable clusters of cells. We show that strong patterns in single cell data will remain despite small perturbations. The results, that are validated based on well-known metrics, show that using this approach yields improvement in correctly identifying the cell types, compared to other existing methods.

AUTOBIOGRAPHICAL STATEMENT

Education

- Ph.D. Computer Science, Wayne State University, Detroit MI, USA, Expected 2020.
- M.S. Information Technology, Iran University of Science & Technology, Tehran, Iran, July 2010.
- B.S. Software Engineering, Alzahra University, Tehran, Iran, October 2006.

Publications

- **Azam Peyvandipour**, Adib Shafi, Nafiseh Saberian, and Sorin Draghici, “Identification of cell types from single cell data using stable clustering", *Scientific Reports* 10, no. 1 (2020).
- **Azam Peyvandipour**, Nafiseh Saberian, Adib Shafi, Michele Donato, and Sorin Draghici, “A novel computational approach for drug repurposing using systems biology", *Bioinformatics* 34, no. 16 (2018): 2817-2825.
- Nafiseh Saberian, **Azam Peyvandipour**, Michele Donato, Sahar Ansari, and Sorin Draghici, “A new computational drug repurposing method using established disease–drug pair knowledge”, *Bioinformatics* 35, no. 19 (2019): 3672-3678.
- Nafiseh Saberian, Adib Shafi, **Azam Peyvandipour**, and Sorin Draghici, “MAGPEL: an autoMated pipeline for inferring vAriant-driven Gene PanEls from the full-length biomedical Literature”, *Scientific Reports* 10, no. 1 (2020): 1-11.
- Adib Shafi, Tin Nguyen, **Azam Peyvandipour**, and Sorin Draghici, “GSMA: an approach to identify robust global and test Gene Signatures using Meta-Analysis”, *Bioinformatics* 36, no. 2 (2020): 487-495.
- Adib Shafi, Tin Nguyen, **Azam Peyvandipour**, Hung Nguyen, and Sorin Draghici, “A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signature”, *Frontiers in Genetics* 10 (2019): 159.