EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR PATIENT STRATIFICATION AND DRUG REPOSITIONING

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Zainab Al-Taie

Dr. Chi-Ren Shyu, Dissertation Supervisor

December 2021

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR PATIENT STRATIFICATION AND DRUG REPOSITIONING

Presented by Zainab Al-Taie,

a candidate for the degree of Doctor of Philosophy, and hereby certify that, in their opinion, it is worthy of acceptance.

_____

Dr. Chi-Ren Shyu

_____

Dr. Rene Cortese

_____

Dr. Mark Hannink

_____

Dr. Jussuf Kaifi

_____

Dr. Jonathan Mitchem

DEDICATION

To

My altruistic mother (Najat Al-Hassani)

For all the sacrifices you have made to raise my siblings and me right, for supporting me in all my decisions, and for the immense trust, care, and love that made me who I am now. Thank you for everything. There are not enough words to describe how grateful I am for everything you have done for us.

My amazing uncle (Emad Al-Hassani)

For being the best uncle who is more like a father in my life. Thank you for taking the responsibility to raise my siblings and me. I am sincerely grateful for all your constant support that helped me to be where I am now. Thank you for your encouragement, understanding, trust, and endless love.

My beloved brother (Ali Al-Taie)

For all the support, caring, and encouragement to overcome all the difficulties I have faced in my Ph.D. and in my life. Thank you for always being there for me through all my ups and downs. I am lucky to have a brother and a friend with a warm heart like you.

My sweet sister (Fatimah Al-Taie)

For your optimism that shines in our lives. Thank you for being my best friend and sister and always being there to listen to me when I need to talk, and this has always helped me have less stress during my journey in the Ph.D. study.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Chi-Ren Shyu, for his continuous support and encouragement, patience and immense knowledge, and for his dedication and enthusiasm for research. His vision, guidance, and trust have enabled me to find my way as an independent researcher. I am deeply grateful to have him as my advisor and mentor.

Also, I would like to thank my committee members, Dr. Rene Cortese, Prof. Mark Hannink, Dr. Jussuf Kaifi, and Dr. Jonathan Mitchem, for always finding the time in their busy schedules to meet and discuss my research. I am grateful for all their insightful comments and encouragement. Additionally, I would like to thank Dr. Christos Papageorgiou for his contribution to my research.

I would also like to thank my current and former colleagues, notably Danlu Liu, Dr. Yan Zhuang, and Dr. Yuanyuan Shen from the Interdisciplinary Data Analytics and Search (iDAS) Lab for all their suggestions and encouragement. I also want to thank Mr. Robert Sanders and Ms. Tracy Pickens for their kindness and professional assistance over the past years.

Finally, I would like to thank my dear friends, foremost among whom are Dr. Amal Al-Yasiri, Ann Huber, Danlu Liu, Iuliia Innokenteva, Dr. Nattapon Thanintorn, and Dr. Yuanyuan Shen, for always being there for me through my ups and downs. Final thanks go to my funding sources from the Informatics and Data Science Research Initiatives of the University of Missouri.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

ABSTRACT

Enabling precision medicine requires developing robust patient stratification methods as well as drugs tailored to homogeneous subgroups of patients from a heterogeneous population. Developing de novo drugs is expensive and time consuming with an ultimately low FDA approval rate. These limitations make developing new drugs for a small portion of a disease population unfeasible. Therefore, drug repositioning is an essential alternative for developing new drugs for a disease subpopulation. There is a crucial need to develop data-driven approaches that find druggable homogeneous subgroups within the disease population and reposition the drugs for these subgroups. In this study, we developed an explainable AI approach for patient stratification and drug repositioning. Exploratory mining mimicking the trial recruitment process as well as network analysis were used to discover homogeneous subgroups within a disease population. For each subgroup, a biomedical network analysis was done to find the drugs that are most relevant to a given subgroup of patients. The set of candidate drugs for each subgroup was ranked using an aggregated drug score assigned to each drug. The method represents a human-in-the-loop framework, where medical experts use data-driven results to generate hypotheses and obtain insights into potential therapeutic candidates for patients who belong to a subgroup. To examine the validity of our method, we implemented our method on individual cancer types and on pan-cancer data to consider the inter- and intra-heterogeneity within a cancer type and among cancer types. Patients' phenotypic and genotypic data was utilized with a heterogeneous knowledge base because it gives a multi-view perspective for finding new indications for drugs outside of their original use. Our analysis of the top candidate drugs for the subgroups showed that most of these drugs are FDA-approved drugs for cancer, and others are non-cancer related, but have the potential to be repurposed for cancer. We have discovered novel cancer-related mechanisms that these drugs can target in different cancer types to reduce cancer treatment costs and improve patient survival. Further wet lab experiments to validate these findings are required prior to initiating clinical trials using these repurposed therapies.

Chapter 1
Introduction

1.1 INTRODUCTION

Patients with the same disease have different reactions to the same drug. This indicates that tailoring drugs to a patient or a group of patients who share common genotypic and phenotypic feature is essential to implementing precision medicine in our healthcare system. De novo drug discovery is a time-consuming, high-cost, and high-risk process. Developing and implementing a new drug can take anywhere between 10-15 years while costing roughly $1.6 billion. The success rate for new drug development is about 2%, with approval rates by the Food and Drug Administration (FDA) declining since 1995 [1, 2]. This highlights the necessity of drug repositioning (DR), or the ability to reposition existing FDA-approved therapeutics for the treatment of additional diseases [3]. DR takes advantage of existing drug therapies already in use and/or at the approval stage to be declared safe for human administration by the FDA [4]. DR reduces the time, cost, and risk associated with the developmental phases of a new drug application, or (N.D.A.), and represents an important strategy for improving patient care.

1.2 SUBGROUPING AND DRUG REPOSITIONING

Patient stratification into subgroups is a crucial step toward applying precision medicine. As we move to a wider implementation of precision medicine and N-of-1 trials in our healthcare system [5], it becomes necessary to move drug discovery in a more patient-centric direction. However, significant barriers exist for the development of new drugs for a small proportion of patients due to excessive development costs and financial burden to

patients. Therefore, DR represents an essential alternative strategy for developing new drugs for patient subpopulations. As these subgroups are identified, we can more specifically align patients and medications to achieve precision-based therapy.

Drug-repositioning methodologies involve both computational and experimental techniques. Computational algorithms represent a significant opportunity for the systematic screening and identification of new indications for existing drugs [6-8]. A majority of the computational analysis components can be grouped into three categories: machine learning, network analysis, and neurolinguistics and language semantics [9, 10]. Drug repositioning using these methods has been undertaken using disease-centric approaches, drug-centric approaches, or combinations of both [11]. In disease-centric approaches, a drug developed for one disease is suggested for another disease after clustering diseases by phenotypic similarity, molecular signatures, and genetic variation [12-15]. Drug-centric approaches accomplish repositioning based on the similarity of drug molecular activity [16, 17]. Some methods are a combination of these approaches based on building drug-drug and disease-disease similarity networks. They then assign drugs based on a meta-path score, predicting disease-drug association [18-20], or the correlation between the gene expression profile of a disease and the genes impacted by a drug [21]. Other methods reposition drugs based on mutations, the expression profile of genes, and protein interactions in the diseases of interest [22]. These methods deal with the broad picture of directing a drug to a new disease, but miss the details represented by the response to these drugs on a subpopulation level. The fact that people with the same disease experience different responses to the same drug highlights the importance of looking more deeply into the details of patient subgroup regimens.

In this study, a novel patient subgroup stratification and drug repositioning method was developed by strategically searching a combinatorial phenotypic space with significant genotypic patterns using a biomedical DR knowledge base [23]. This is a unique network-based and explainable data mining computational approach for subgroups discovery and drug repositioning. A heterogeneous knowledge base was adopted from the 'hetionet' project to create our DR knowledge base (DR-KB) [24]. Patients' phenotypic and genotypic data was utilized in conjunction with the heterogeneous knowledge base to provide the most accurate depiction of living systems and their complexity. This heterogeneity gives a multi-view perspective to find new indications for drugs outside of their original use. This DR approach represents an effective, applicable, and significant opportunity to approach precision medicine using an explainable and data-driven computational method. The Cancer Genome Atlas (TCGA) was used for the case studies as it contains a large volume of clinical, pathologic, and molecular features which allow us to create highly granular patient subgroup DR recommendations.

1.3 RELATED WORK

The importance of repositioning drugs to tailor treatment for homogeneous subgroups within a heterogeneous disease population has been demonstrated by previous studies. Most of the existing methods to stratify patients into disease subcategories are based on clinicopathologic features, with some malignancies having shifted towards molecular subtypes [25-28]. The primary method has been to identify drug repositioning candidates for subgroups of patients based on targeting particular genes proven to have a role in disease development. Gouravan, et al. [29] demonstrated that drugs could be repositioned for subgroups of sarcoma patients with well-known mutations that frequently occur, such as

finding candidate drugs targeting a BRAF mutation. Simon, et al. [30] focused on a mutation in the RUNX1 gene and studied drug sensitivity to identify candidate drugs for repositioning in patients with acute myeloid leukemia (AML) and a mutation in this gene. Yoshida, et al. [31] produced studies that focus on Myc mutation and investigated this gene family's therapeutic potential across different cancer types. Another method is stratification based on known genotypic variations. After identification, the critical genotypic characteristics of each subgroup are used to identify drugs and targets for repositioning. An example is the repurposing of subtype-specific drugs for breast cancer after the identification of three different modules of Triple-negative breast cancer (TNBC) based on protein-protein interaction networks [27]. Nepal, et al. [32] stratified Intrahepatic cholangiocarcinoma (iCCA) patients based on mutations in three classifier genes, IDH, KRAS, and TP53, and studied their ability to induce substantial downstream molecular heterogeneity and pharmacogenomic potential.

Patient stratification and drug repositioning has also been achieved by clustering patients based on a set of gene mutations. Lind and Anderson [33] used machine learning to predict the activity of small-molecule drugs against cancer cells using mutations in oncogenes. In their study, patients were clustered based on their mutation profiles in specific oncogenes and drugs selected targeting these mutations. Gligorijević, et al. [34] also applied machine learning methods to identify patients based on mutation and drug data along with molecular interactions to reposition drugs based on targets in each patient cluster. Additionally, data mining has been used to stratify patients based on molecular features and to prioritize drug targets for repositioning within pre-defined molecular subgroups. Chen

and Xu [26] developed a computational method to repurpose drugs for glioblastoma molecular subtypes using human cancer genomics combined with mouse phenotype data.

Though these methods represent a significant step in a promising direction, discovering drugs for a large number of subgroups of patients requires a more comprehensive exploratory approach where multiple factors are considered to address the challenge of patient diversity. The methods outlined above, among others, have shown the importance of patient stratification and have produced promising results. However, patient stratification and DR strategies focused on subgroups with commonly known genotypic characteristics may miss the importance of phenotypic characteristics during the stratification process. Using these methods to solve biomedical problems that impact human life required the implementation of the Explainable AI (XAI) concept, where the system offers humans the ability to analyze and understand its action and the reasons behind any prediction in order to overcome the black-box challenge of AI in medicine [35]. The Explainable AI concept focuses on transparency, meaning that the actions that affect human life should be explained in a format that is understandable by humans and should show the underlying phenomena of any prediction [36]. The importance of developing explainable computer-based systems to solve biomedical problems has been expressed in studies that include prediction based on medical image processing and genomics analysis [37, 38]. Some studies addressed the need for explainable biomedical systems by building interactive ML (iML) models [39, 40]. The goal of iML is to enable algorithms to explain each step to users and enable them to correct the provided explanation [41]. Still, the explanations provided by these methods require further study [42]. For patient stratification and drug repositioning, the explainability of machine learning and data mining results can be improved by including

the ability to provide insight regarding the underlying biological mechanism that is unique to a subgroup as well as how the perturbation of biological entities contributes to drug selection for a given subgroup.

The method in this study aims to build an explainable AI system using data mining and network analysis. This is a step toward building advanced Explainable AI systems that imitate the human cognitive system in which humans make sense of the world by recognizing patterns. This method provides an exploratory stratification process through the investigation of not only the phenotypic inclusion and exclusion criteria, but also the genotypic characteristics of subgroups that differentiate them from the larger disease population, as well as druggability based on these characteristics. Moreover, this method provides the flexibility to stratify a patient to multiple subgroups. This gives medical practitioners the ability to consider alternative treatments that remain specific for each patient.

Chapter 2
Methods

2.1 INTRODUCTION

This Patient Stratification and Drug Repositioning (PSDR) framework is composed of three modules. (1) Materials and Data Preprocessing: Patients' genotypic and phenotypic data was preprocessed and categorized to be the input to the patient statification algorithm. In this module, a heterogeneous KB was integrated with patient data (Section 2.2). (2) Subpopulation Discovery: An explainable AI method for drug repositioning and subgroup discovery of a disease population was developed. For each subgroup, a heterogeneous network was created based on the phenotypic characteristics and gene signatures (Section 2.3). (3) Drug Candidate Evaluation: For each subgroup resulting from the subpopulation discovery module, a drug score was calculated within each network and ranked for prioritization and recommendation of repositioning for each identified subgroup (Section 2.4).

2.2 MATERIALS AND DATA PROCESSING

The input data for the patient stratification framework consists of genotypic and phenotypic variables for a disease population. The phenotypic, genotypic, and heterogeneous biomedical networks are used to guide subgroup discovery and recommend drugs for these subgroups (Figure 1-Module 1). In this study, the genotypic and phenotypic data for patients were obtained from TCGA. As part of the human-in-the-loop process, a physician panel involved in the care of cancer patients selected the phenotypic and clinical variables to be included in the analysis. Additionally, many of these phenotypic variables were continuous, which required stratification into categories for inclusion in the data mining algorithm. The medical guidelines and the physician panel guided the categorization of all continuous variables. For example, the original data set of colorecata cancer (CRC)

contains the age of each patient. Patient age was categorized into three age groups, which are <50, 50-69, and >69. The genotypic data in this study are genes differentially expressed between normal and tumor tissues. The differential expression analysis using edgeR was implemented on the RNA-seq data of the patients. The dimensionality reduction was made by deciding the p-value to be less than 0.05. In addition, a neo4j graph representation of different biomedical entities and the relations between them was used as our DR knowledge base (DR-KB), which contains 11 different types of biomedical variables (Gene, Biological process, Cellular Component, Molecular function, Pathway, Anatomy, Drug (Compound), Side effect, Pharmacological class, Disease, and Symptom) from hetionet [24].

2.3 SUBPOPULATION DISCOVERY

In this section, the stratification process is described based on clinical and genomic characteristics to enable drugs to be directed to a selected list of homogeneous groups within the heterogeneous disease population. This section has been published in the Journal of Biomedical Informatics [43].

2.3.1 PATIENT STRATIFICATION

Finding a homogeneous subgroup cohort is a crucial step in enabling precision medicine. In this study, the focus was to systematically and strategically group patients into phenotypic subgroups based on their genotypic characteristics. The exploratory data mining [23] method was extended by integrating network analysis to guide the subpopulation discovery process. This exploratory data mining method provides an automatic subpopulation discovery tool that computationally investigates a large pool of subpopulations that have underlying factors differentiating each subpopulation within a given larger group.

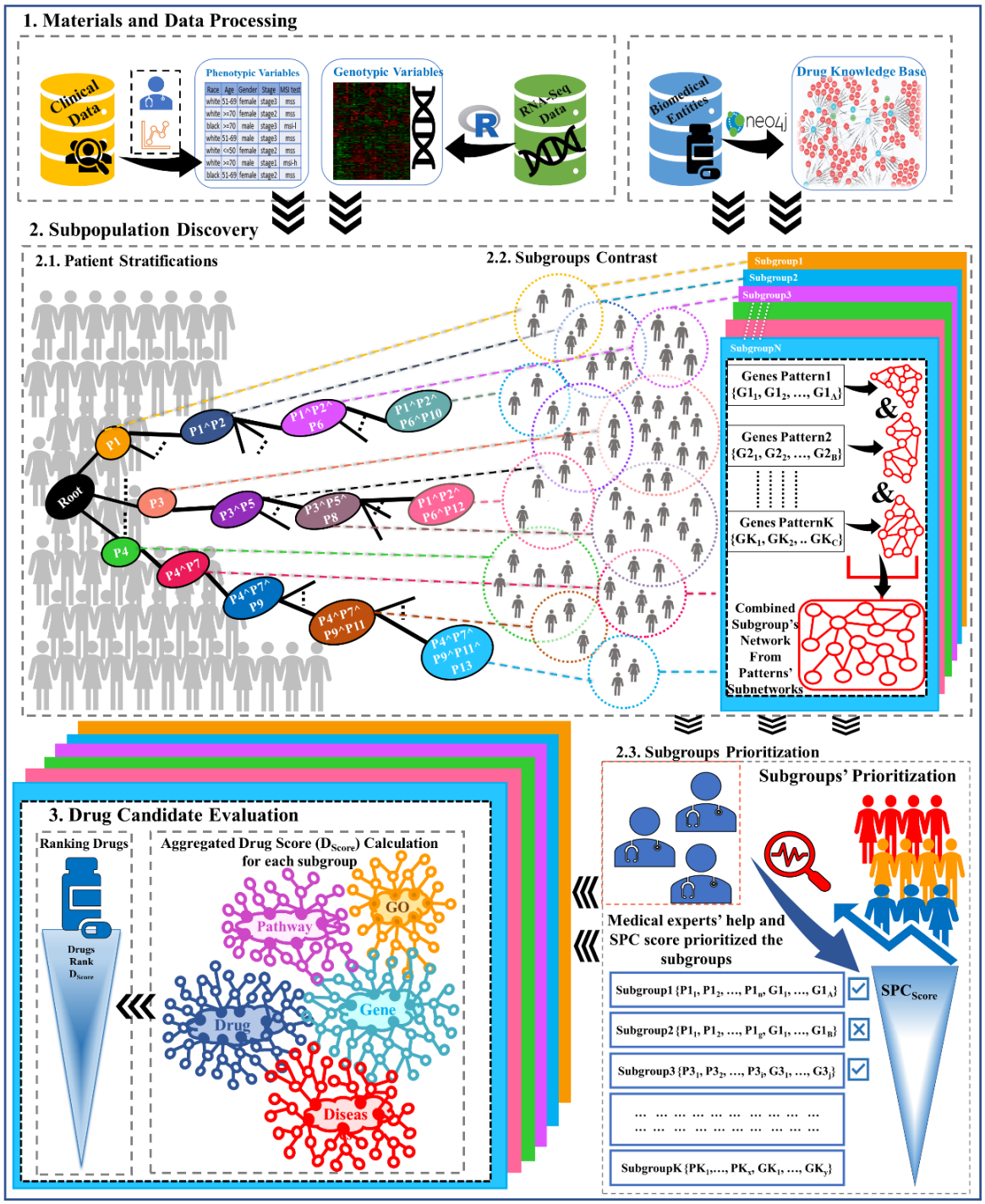

Figure 1: Patient Stratification and Drug Repositioning Framework. Module 1: The input data consists of patient phenotypic and genotypic characteristics or variables and the DR-KB. Module 2: The subpopulation discovery and evaluation process. Module 2 consists of 3 submodules. Module 2.1: During the path expansion process in which we add or delete a node, module 2.2 is applied to evaluate the contrast and identify the significance of adding or removing a node. Module 2.2: This module is used to calculate the contrast score for each candidate subgroup with the outer population by applying contrast pattern mining and network differentiation. Module 2.3: The subpopulation contrast score (SPC_{score}) is used to rank the candidate subgroups. Medical experts conduct further evaluation of the subgroups that have more potential in clinical settings. Module 3: It is for drug evaluation. In this module, an aggregated drug score (D_{score}) is calculated for each drug in each subgroup network that is created from all the contrast pattern subnetworks. This score is used to rank the drugs for each subgroup to connect the most relevant drugs to a given subgroup.

The results are presented as subgroups, each defined by a set of population criteria and underlying factors which differentiate each subgroup from the entire population. These criteria are the phenotypic variables, such as gender, age, and cancer stage. An example subgroup could be males aged less than 50 years with stage II cancer. When a phenotypic feature is added, a focus subpopulation is created and contrasted with the rest of the population. Adding additional population variables is desired, assuming there is statistical evidence to do so. The determination of the significance of a subgroup is based on underlying factors which are the genotypic patterns that are statistically unique to the subgroup in comparison to the rest of the population by utilizing Contrast Pattern Mining (CPM) [44].

The patient subgroup stratification module takes a three-level approach. The top-level method, path expansion, includes a large number of second-level floating subgroup selection processes, each of which is supported by a series of third-level Inclusion and Exclusion procedures. This method is exploratory and differs from a decision tree approach in which samples are divided based on the decision for each node, and each leaf node contains a group of samples which are exclusively in a particular node. Unlike a traditional decision tree, the proposed method has a large number of dynamic fanouts for each node without dividing the samples during the expansion process, and each node represents a subgroup. As a result of the patient subgroup stratification process, a patient could be in multiple subgroups through branching expansion.

For example, female patients could be grouped into (female, stage III) and (female, age<50) which potentially contain patients in both groups. To better understand the complexity of this method, let n_p be the number of phenotypic (population) variables with

an average n_c categories per variable. There are ${n_p}^{n_c}$ subgroups in the exploratory space, which is an unmanageable scale. Therefore, the core of the patient subgroup stratification module (Figure 1-Module 2) is to automatically and efficiently identify a large number of viable patient subgroups using phenotypic variables, where unique and qualifiable genotypic characteristics are shared by the majority of the patients in the subgroup. This approach differs from traditional greedy algorithms in two ways: (1) path expansion selects top potential subgroups that have equivalent performance in druggability of the best and local optimal selection at any stage of the process to ensure the broadness of selected subgroups that are equally viable, and (2) a series of inclusion and exclusion criteria of phenotypic variables using the floating selection approach [45] are performed to avoid simple greedy selection of subgroups.

For each path in Figure 1- Module 2.1 and Figure 2, the algorithm begins by choosing a single phenotypic variable ($P_i = C_i$) as a base subgroup with the most significant contrast against the remainder of the population. The contrast is measured based on the genotypic patterns and subgroup network differentiation from the outer population. Genotypic results guide the algorithm to perform the next inclusion or exclusion of population variables. The subgroup is identified in the stratification process as a group of patients who have the same phenotypic features and share common genotypic patterns and network perturbation patterns, which are unique to that subgroup as compared to the rest of the population.

The contrast calculation will be explained in detail in Section 2.3.2. During the floating subgroup selection process, the inclusion step (INCLUSION() function in the pseudo-code and SG2 and SG3 in Figure 2) adds a new phenotypic variable $P_j = C_j$ to the

previous subgroup to generate a more focused subgroup ($P_i = C_i \hat{} P_j = C_j$). After each inclusion step, the exclusion step (EXCLUSION() function in the pseudo-code and SG5 and SG9 in Figure 2) is adopted to exclude a less significant move made previously by removing a variable from the "greedy" subgroup if the result of the exclusion process has better performance. For example, when the subgroup is ($P_i = C_i \hat{} P_j = C_j \hat{} P_k = C_k$) at the third inclusion step, the exclusion step will remove the previous less significant move ($P_i = C_i$) from the current subgroup if the newly generated subgroup ($P_j = C_j \hat{} P_k = C_k$) has better "druggability gain" (Figure 2, 2nd path in the red expansion process). The druggability measurement in this work is computed by the potential drug targets using unique genotypic patterns in the current subgroup contrast with the remainder of the population. The patterns of genotypic variables are extracted using the PATTERN_MINING() function, where the algorithm selects the genotypic patterns that frequent in the most patients in a given subgroup. The knowledge base is queried to create a heterogenous network of biomedical entities that interact with these patterns including the protein-protein interactions using the NETWORK_CREATION() function.

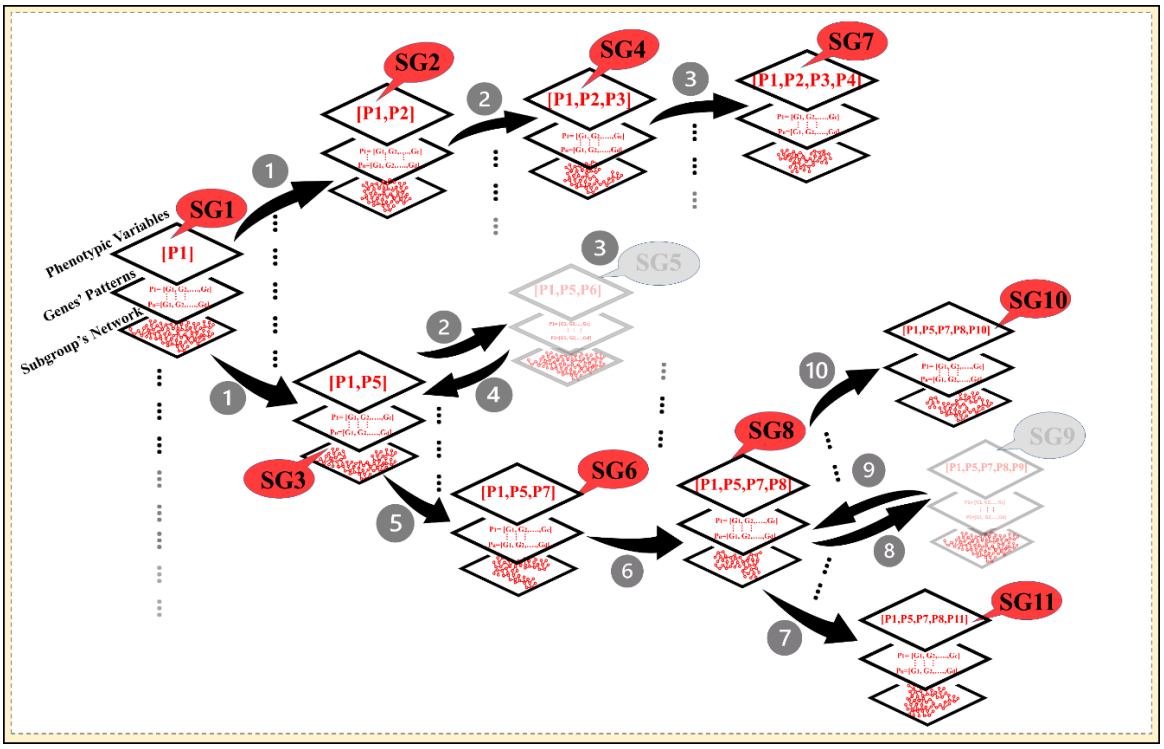

Figure 2: The subpopulation discovery and evaluation process. This is the floating and path expansion process, where we have different starting points on different computational nodes. For each point, we add or delete a node based on the contrast score. The contrast is evaluated by identifying the significance of adding or removing a node. The contrast score (SPC) is calculated for each candidate subgroup as compared to the outer population by applying contrast pattern mining and network differentiation. Each point represents a potential subgroup, and three layers represent it. The first layer is the phenotypic variables. The second layer is the genes' pattern that are frequent in that group of patients. The third layer is the biomedical interaction of the patterns in the 2nd layer after mapping them to the DR-KB to create the subgroup's network. The SPC score is calculated after comparing the subgroup's network with the rest of the population.

Algorithm: Subgroups Discovery and Drug Repositioning

Inputs:

P(D): The phenotypic variable set for dataset D.

SPC(k): Contrast score for subgroups with k variables.

α: Stopping criteria.

M=maximum number of phenotypic variables for contrast subgroups.

Output:

Resulting subgroup with highest contrast *SGI*

Recommended drugs for *SGI*

Start

1: $SGI \leftarrow \emptyset$; $k \leftarrow 0$; $SPC(k) \leftarrow 0$;

2: WHILE $k < 2$ DO:

3: Inclusion (P(*D*), *SCGI*)

4: $k \leftarrow k + 1$

5: END

6: While ((SPC (*k*) – SPC (*k*-1))/SPC(*k*)) > α AND $k < M$ DO:

7: $P_{inclusion}$ =INCLUSION (P(*D*), *SGI*)

8: $P_{exclusion}$ =EXCLUSON (P(*D*), *SGI*)

9: IF ($P_{inclusion} = P_{exclusion}$) THEN

10: $k \leftarrow k + 1$

11: $SPC(k) \leftarrow SPC(SGI)$

12: ELSE

13: GO TO LINE # 8

14: END

15: DRUG_REPOSITIONING(*SGI*)

End

<u>Function: INCLUSION (P(D), SGI)</u>

1: *SGS*: potential subgroup set.

2: $SGS \leftarrow \emptyset$

3: FOREACH phenotypic variable $P_i \in P(D)$ DO

4: $CP_{Set}(P_i) \leftarrow [Pairs(\forall\ categorical\ value(P_i) \leftrightarrow outer\ population)]$

5: FOREACH CP($C_{i,a}$, _) ∈ $\mathrm{CP_{Set}}(P_i)$

6: $SGI_{temp} \leftarrow SGI + \mathrm{CP}(C_{i,a}, \_)$

7: $SG_1 \leftarrow D(C_{i,a})$

8: $SG_2 \leftarrow D - D(C_{i,a})$

9: $PTR_1 \leftarrow PATTERN\_MINING(SG_1)$

10: $PTR_2 \leftarrow PATTERN\_MINING(SG_2)$

11: $NW_1 \leftarrow NETWORK\_CREATION(PTR_1)$

12: $NW_2 \leftarrow NETWORK\_CREATION(PTR_2)$

13: SPC(SGI_{temp}) ← $CONTRAST\_CALCULATION(NW_1, NW_2)$

14: Add SGI_{temp} to *SGS*

15: END

16: END

17: $SGI_{highest} \leftarrow The\ highest\ SPC\ subgroup$

18: $SGI \leftarrow SGI_{highest}$

19: Remove phenotypic variables of $SGI_{highest}$ from P(*D*)

Function: EXCLUSION (P(D), SGI)

1: *SGS*: potential subgroup set.

2: $SGS \leftarrow \emptyset$

3: FOREACH CP($C_{i,a}$, _) ∈ $CP_{Set}(P_i)$

4: $SGI_{temp} \leftarrow SGI - \mathrm{CP}(C_{i,a}, \_)$

5: $SG_1 \leftarrow D(C_{i,a})$

8: $SG_2 \leftarrow D - D(C_{i,a})$

9: $PTR_1 \leftarrow PATTERN\_MINING(SG_1)$

10: $PTR_2 \leftarrow PATTERN\_MINING(SG_2)$

11: $NW_1 \leftarrow NETWORK\_CREATION(PTR_1)$

12: $NW_2 \leftarrow NETWORK\_CREATION(PTR_2)$

13: SPC(SGI_{temp}) $\leftarrow CONTRAST\_CALCULATION(NW_1, NW_2)$

14: Add SGI_{temp} to *SGS*

15: END

17: $SGI_{highest} \leftarrow$ *The highest SPC subgroup*

18: *SGI* $\leftarrow SGI_{highest}$

19: Add phenotypic variables of $SGI_{highest}$ back to P(*D*)

In this work, instead of tracking only one path, the path expansion process considers the top β% paths based on the druggability gain measurements as the potential successful subgroups at each inclusion and exclusion step, where β is defined as a tracing factor that expands the search to top performers. For example, in parallel to the selection of *P1* at the root of the tree structure in Figure 2, there are multiple paths that are among the top β% in druggability performance. This fanout number could range between 1 and $0.1 * \sum_{p=1}^{n_p} C_p$ branches per node, where C_p is the number of categorical values of phenotypic variable p. The evaluation of the druggability gain at each step is done by applying subgroups contrast (Section 2.3.2). Unlike traditional optimization methods that result in sub-optimal solutions, this deep mining process will generate a sizable number of subgroups through the expansion

process and less greedy results through the floating process. This process is deep mining because the algorithm does not simply offer a single model through a traditional greedy approach. Rather, it takes a sizable number of branches during each step and the final export often generates hundreds to thousands of subgroups. It is analog to deep learning methods to scale the number of encoding and decoding layers with a massive number of neurons that cannot be trained without today's computation power. This method works in a deep and wide manner to find subgroups. It is "deep" because the algorithm proceeds from a most general subgroup to a more specific subgroup for each path. This happens by going deeper in each path. The exploratory search will be terminated in each path when the algorithm gets into a most highly specific subgroup with the highest contrast score that cannot be improved further. This algorithm works in a "wide" manner because it explores a large number of "equally creditable" paths from each stratification decision to identify new subgroups.

The subgroup prioritization method (Section 2.3.3) is used to decide whether to keep a node (feature) or remove it if the parent node has more significance than the child node, meaning that the child node does not add further specificity, thus ending that path. A distributed computing framework with Apache Spark was utilized to run this computationally expensive process. The Big O for the algorithm is $O(\left(\beta n_c {n_p}^2\right)^{n_c})$, where β is the tracing factor, n_p is the number of phenotypic (population) variables, and n_c is the average number of categories per variable. After finishing the floating and expansions, the candidate subgroups are prioritized using an index that evaluates the aggregated contributions of all the extracted contrast patterns within each subgroup based on the number of contrast patterns (e.g., co-occurring mutated genes) and the significance (e.g.,

druggabilities) of those patterns. The final output of this tool is a ranked subgroup list. For each subgroup, the contrast patterns which differentiate a given subgroup from the entire patient population were provided. These patterns present insight into underlying differences among subgroups and are valuable for further study or for clinical trials.

2.3.2 SUBGROUP CONTRAST

As discussed in the previous section, the evaluation of subgroup significance is performed by measuring the contrast between the subgroup and its outer population. For each candidate subgroup representing a set of phenotypic characteristics, the algorithm finds all genotypic patterns that are frequent within the subgroup but infrequent in the remaining population. Support [46] is used to evaluate whether a given pattern is frequent in a subgroup and growth rate [44] to evaluate the contrast of the pattern in the selected subgroup. In addition, each pattern is evaluated based on its druggability. By mapping these patterns to the DR-KB, the contrast of each subgroup with the outer population is evaluated using multiple biomedical entities, including gene, biological process, cellular component, molecular function, pathway, anatomy, side effect, pharmacological class, disease, symptom, and drugs that are connected to these patterns in the DR-KB network. Because there are multiple patterns in each subgroup, an overall evaluation of the subgroup can be assessed by aggregating the contributions of the selected patterns. This subgroup evaluation is used to assess the druggability gains on the floating and expansion process in Section 2.3.1.

Let D be the patient dataset in a subgroup, which includes n genotypic variables, $G = (g_1, g_2, \ldots, g_n)$. Pattern p that is commonly shared within patients in a given subgroup is defined as a set of genotypic variables, such as $p = (g_{1,e1}, g_{2,e2}, \ldots, g_{i,ei})$, where $g_{i,ei}$ is the

expression level or mutation status of gene i. The expression level or the mutation status should be represented as a categorial value. This process is accomplished using the PATTERN_MINING() function in the pseudo-code.

The pattern is "frequent" if its support is greater than a user-defined threshold. The support of pattern p is the number of records (patients) that have that pattern ($|<D,p>|$) divided by the total number of records in the dataset D ($|D|$):

$$Support(p, D) = \frac{|<D,p>|}{|D|} \quad (1)$$

To find the contrast pattern (cp) between the focus subpopulation and the rest of the population, S_{G1} represents the focused subgroup and S_{G2} represents the remaining population, where $S_{G2}=D-S_{G1.}$ The support of the contrast pattern should be significantly different between S_{G1} and S_{G2}. Let s_1 be the support of a contrast pattern in S_{G1} and s_2 the support of the same pattern in S_{G2}. The growth is used to measure the difference between the two groups. The growth of contrast pattern cp between subgroup S_{G1} and the remaining population S_{G2} is defined as follows:

$$\text{Growth}(cp, S_{G1}, S_{G2}) = \frac{Max\{s_1,s_2\}}{Min\{s_1,s_2\}} \quad (2)$$

The growth ratio is normalized to be between 0 and 1 using an extended version of the tanh function [47]. Let α be the threshold for the support and β the threshold of growth rate. To ensure that a cp is frequent and has significant differences between the two groups, the following condition should be held:

$$(\text{Support}(cp, S_{G1}) \geq \alpha \text{ OR Support}(cp, S_{G2}) \geq \alpha) \text{ AND } (\text{Growth}(cp, S_{G1}, S_{G2}) \geq \beta) \quad (3)$$

This condition identifies two sets of contrast patterns CP_1 and CP_2 for the target subgroup and the outer population, respectively. For each contrast pattern cp_n with multiple genotype variables, the subset of the pattern $cp_i \subseteq cp_n$ will be kept when Growth(cp_i, S_{G1}, S_{G2})−Growth(cp_n, S_{G1}, S_{G2}) > 0. These selected contrast patterns are utilized to evaluate each subgroup during the floating and path expansion procedure discussed in Section 2.3.1.

For the purpose of drug repositioning, contrast patterns should embed the druggability of the candidate subgroups. For a gene set in a pattern that is frequent in the focus subgroup but not in the remaining population, the DR-KB is queried to extract the biomedical entities connected to each gene in the pattern. Each pattern is represented by a subnetwork of the DR-KB. An aggregated network for a given subgroup is obtained by integrating all frequent patterns. To measure the significance of the subgroup based on its relevant patterns and druggability, a contrast score (SPC_{Score}) is calculated using the CONTRAST_CALCULATION() function in the pseudo-code. The calculation of this score is based on values of two major components that were multiplied to obtain the contrast score (Equation 5).

The first component of the product in Equation 5 measures the contrast of the given subgroup based on the genotypic characteristics of the patients within that subgroup, while the second component measures the contrast of the given subgroup in comparison to the outer population on different levels of biomedical entities that are unique to the subgroup. In the first component, T is a parameter related to the population size t, where $T = \{t, 1/t\}$. $T = t$ when a large population is preferred and $T = 1/t$ when a smaller population is preferred, such as a study of a rare disease. M is the average population size of randomly chosen contrast subgroups prior to path expansion. J_{org} and J_{avg} are calculated based on the

J-value, that is a quantitative index to evaluate the overall quality of a set of contrast patterns in the subgroup. The *J* value for each subgroup is used to prioritize it among all discovered subgroups. This *J* value measurement was inspired by the g-index, which is commonly used to evaluate the productivity of a scholar. If a subgroup (scholar) has a set of patterns (articles), the J-index (g-index) is measured by ranking them in decreasing order based on their growth rate (citations) and then by taking the largest number such that the top J contrast patterns (top g articles) have cumulatively received at least J^2 (g^2) scores. The J-value is defined as follows:

$$J^2 \leq \sum_{i \leq J} Growth(cp, SG_1, SG_2) \quad (4)$$

In the second component of Equation 5, different biomedical entities are considered in addition to patient genotype patterns that are unique in the subgroup of interest. The biomedical entities that are unique to the subgroup but not to the entire disease population are also considered. This is the motivation to include the second component of the product in the equation.

$$SPC_{Score} = [(T * J_{org} + M * J_{avg})/(T + M)] * [1 - (\sum_{i=1}^{n}((E_{i,1} \cap E_{i,2}) / (E_{i,1} \cup E_{i,2}))/n)] \quad (5)$$

Patient stratification is accomplished by considering the patients' specific data and a comprehensive biomedical knowledge base. The knowledge base is heterogeneous to represent the different aspects of the human biological system and the prospective effects of drugs on this system. Each component contributes to the biological and druggable meaningfulness for the patient stratification process. Taking all the biomedical similarities and differences into account in determining the subgroups is essential to arriving at a more comprehensive assessment for the subgrouping. To address the heterogeneity of biological

systems in the context of drug repositioning, DR-KB network similarity is integrated into the contrast score calculation.

Let *BioE* be the types of the biomedical entities in the network. *BioE* = {Gene, Biological process, Cellular component, Molecular function, Pathway, Tissue (Anatomy), Drug (Compound), Side effect, Pharmacological class, Disease, and Symptom}. These different biomedical entities are essential for calculating subgroup druggability because the drug effect is not only dependent on the genes as isolated entities. These genes are part of different biomedical entities, and perturbations of these genes have differing impacts through the relationships of and interactions with various biomedical entities. For example, genes with disease [48-50], pathways [51, 52], GO [53-55], tissues [56], side effect [57, 58], and the pharmacological classes [59] were used in drug repositioning. In the knowledge base, there are 11 possible biological entity types in the network ($n = |BioE| = 11$). The interactions between genes are based on protein-protein interaction but are represented by the gene names that encode these proteins to reduce the complexity. Gene–interacts–Gene edges represent the physical interaction of the protein products of these genes [24].

$E_{i,1}$ is biomedical entity type i in the focus group's network, $E_{i,2}$ is biomedical entity type i in outer population network, $E_{i,1} \cap E_{i,2}$ represents the number of common entities in entity's type i between the focus group and outer population. $E_{i,1} \cup E_{i,2}$ represents the number of all possible entities in entity's type i between the focus group and outer population. By dividing the number of common entities by the number of all possible entities, the score of similarity between the two groups is obtained. Subtracting the similarity score from 1 gives the percentage of difference (contrast) between the two groups based on the extracted knowledge from drug repositioning knowledge base. SPC_{Score}, which is a

product of the two components, represents the contrast score between the focus group and the outer population (Figure 1- Module 2.2). This method ensures that each subgroup with common phenotypic characteristics is distinct from the rest of the population. It also ensures homogeneity within each subgroup by including patients who have similar genotypic features. As the method ensures homogeneity within each subgroup, it allows a patient to exist in multiple subgroups to provide desirable flexibility in the healthcare setting. Critically, this provides the ability to find alternative treatment options when the first or second line of treatment fails. Therefore, heterogeneity among subgroups should not be enforced. At the same time, the heterogeneity between the more general subgroup and the more targeted subgroup arises on the genotypic level, where having a smaller subset of the population could enable the algorithm to discover new genotypic patterns that were not statistically significant in a more general population. To ensure statistical significance, we kept only the subgroups with a p-value < 0.05.

2.3.3 SUBGROUPS PRIORITIZATION

The number of candidate subgroups selected by the floating and expansion process could be hundreds. The SPC_{Score} is used to rank the subgroups. The higher the SPC_{Score}, the higher the potential for drug repositioning. Because this method was developed to improve patient care, all steps should be explainable and acceptable for practitioners. For clinically meaningful results, a physician-in-the-loop process was necessary to prioritize the subgroups further using a two-phase method. First, physician-in-the-loop provides a filtering mechanism where the focus will be on only a subset of the subgroups instead of going through the hundreds of subgroups resulting from our method. Second, the physicians may decide the most relevant subgroups by evaluating the top subgroups using the SPC_{Score}

or using initial hypotheses formed by clinical observations and literature to prioritize all candidate subgroups. For example, in the CRC study, the physician investigators chose to focus on the subgroups with microsatellite (MS) status as one of the clinical variables in the CRC case study in which seven subgroups with Microsatellite Instability (MSI) test results were further examined as a phenotypic characteristic among the statistically significant subgroups (p-value<0.05). The rationale for the selection of groups based on MS status is related to therapeutic selection and tumor biology [60]. Microsatellite instable tumors are associated with hypermutation due to the inactivation of mismatch repair genes via either germline mutation or methylation, accounting for 13-15% of CRCs. The remaining 85% of colorectal cancers develop via the chromosomal instability pathway, referred to as microsatellite stable, following a well-described pathway acquiring mutations through the adenoma to carcinoma sequence as described in seminal work by Vogelstein, et al [61]. While these tumors appear to be biologically different, most critically, these tumors are also characterized by different prognosis, response to standard therapy, and response to novel therapy including both targeted and immune-based therapy [62-64]. Therefore, this designation was felt to be highly clinically relevant. The subgroups that are selected through the physician-in-the-loop process are then chosen as the input for the next step in which drug candidates are evaluated and analyzed for each subgroup.

2.4 DRUG REPOSITIONING

A drug repositioning algorithm is used to reposition drugs that fit best for the patients in each subgroup. As mentioned before, a subgroup network consists of heterogeneous entities. In each network, there is a number of entities of each type including drugs. To find drugs with the highest potential to treat the patients in a subgroup, the drugs within each network must be ranked according to their relevance to the genotypic characteristics (Figure 1-Module 3). A drug scoring function was developed which included different factors to put the drug impact within the subgroup network (SGNW) and its impact in general on the human cells into perspective. The development of the scoring system consists of 2 stages. First, an aggregated drug score was developed to rank the drugs within each network based on the number and the weight of direct and indirect conections for each drug within a subgroup's network (Section 2.4.1). Second, the drugs scoring function was upgraded to include wider range of factors to assess the impact of drugs on the cell in general in addition to their impact within the subgroup's network (Section 2.4.2).

2.4.1 DRUG EVALUATION USING AGGREGATED DRUG SCORING

The biomedical entities directly and indirectly connected to each gene, which is DEG, in the patterns are extracted from the DR-KB (Figure 1- Module 2.3), which is represented as a neo4j directed graph $G=(V_G, E_G)$. For drug repositioning, drugs within each subgroup's network must be evaluated based on the connectivities of each drug in the network. This evaluation depends on how many genes in the subgroup are affected by that drug and the connectivity between the genes and other biomedical entities in the network. To accomplish this, an aggregated drug score is calculated over each subgroup's network for each drug.

Let $GS=(g_1, g_2, \ldots, g_n)$ be the genotypic signature of subgroup S_{Gi}. A graph H composed of a collection of sub-networks for all genes in GS can be obtained through the following equation:

$$H = \{(V_H, E_H) \mid (V_H \subseteq V_G) \wedge (E_H \subseteq E_V) \ \forall g \in GS\} \quad (6)$$

, where V_H is a set of vertices in which g is reachable. The resulting network is representative of the subgroup with different biomedical entities, including genes, biological processes, cellular components, molecular functions, pathways, tissues, diseases, and drugs. Due to multiple candidate drugs in each subgroup's network, the drug prioritization method is required to rank drugs within each network. An aggregated weight calculation algorithm is used to prioritize the drugs in each subgroup network.

Let G' be the subgroup network. $G'=(V', E')$ is a rooted graph where all vertices are directed toward the drugs d as shown in Figure 3. G' has edge weights w: $E' \mapsto \mathbb{R}$. The assignment of weights is based on interaction types. For example, gene-gene interaction has a higher weight than that of tissue-disease interaction.

In the subgroup network, only the entities directly or indirectly connected to genes that, in turn, are connected to drugs were retained. A gene is determined to be connected to a drug if:

- There is a direct edge $e_{ij}(v_j,v_i) \in E'$ that connects gene v_j to drug v_i (gene g_1 and drug d_i in Figure 3-A).
- There is an indirect edge, $e_{ij}(v_j,v_i)$, that connects one or more genes of the subgroup's genes set to a drug through another gene if $e_{ix}(v_x,v_i) \in E'$ and $e_{xj}(v_j,v_x) \in E'$, where v_x is a gene (g_a in Figure 3-B).

- There is an indirect edge, $e_{ij}(v_j,v_i)$, that connects a gene to a drug through a disease if $e_{ix}(v_x,v_i) \in E'$ and $e_{xj}(v_j,v_x) \in E'$, where v_x is a disease (s_n in Figure 3-C).

First, a score is calculated for each gene in G'. The score is the weighted sum of all paths that connect a leaf node to the given gene. Let P be a path that goes from a leaf node v_k to the given gene v_i, $P=<v_i, v_1, \ldots, v_k>$, and N_{ik} is the total number of paths directed toward node v_i from any leaf node v_k.

$$w(vi) = \sum_{p=1}^{N_{ik}} \sum_{i=1}^{k} w(v_{p,j}, v_{p,j-1}), \quad (7)$$

, where $w(v_i)$ is the weight or the score of the given gene v_i and $w(v_{p,i-1}, v_{p,i})$ is the weight of the edge that pointed from node v_{i-1} to the node v_j in path P. A drug score (D_{Score}) can be determined by calculating the sum of the weights of all genes that have interactions with that drug.

$$DScore = \sum_{n=1}^{d_{G'}^{-}(d_i)} w(v_n), \quad (8)$$

where $d_{G'}^{-}(d_i)$ is the in-degree of vertex d_i in G'. $w(v_n)$ is the weight of a gene with index n within the the set of genes that are connected to drug d_i.

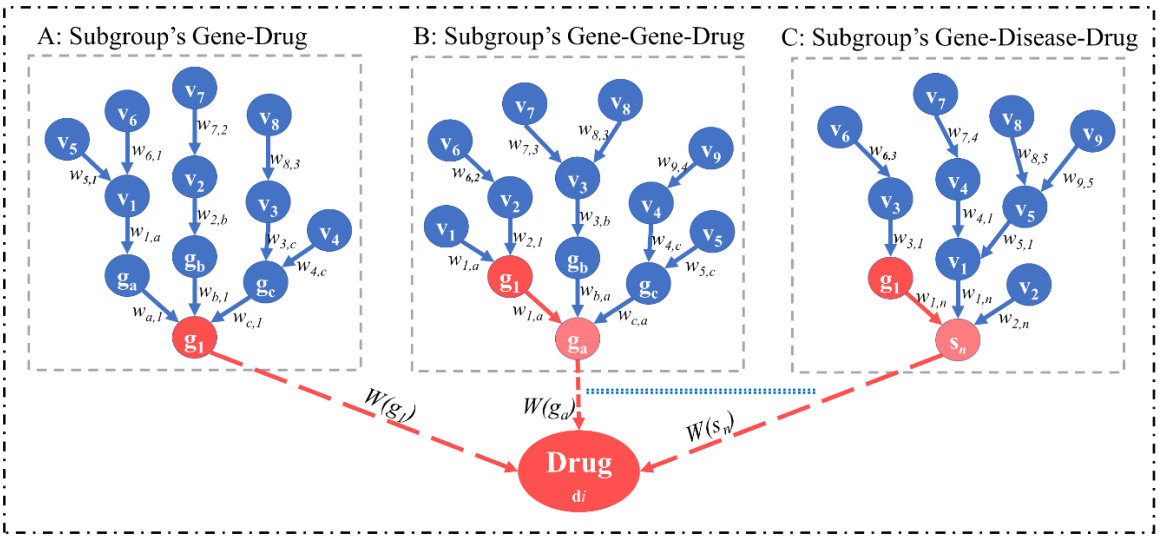

Figure 3: An example of aggregated drug score calculation for each drug, where v's represent the vertices of that subgroup's network, and w's are the weights of the edges that connect one node to another. The aggregated score is calculated layer by layer from the leaf nodes to the root nodes, the genes (g_1, g_a, and s_n). g_1 is a gene from the subgroup's genes set. The final drug score is calculated by summing up the aggregated score (w(g's)) of all the genes connected to that drug.

The subgroups discovery and drug repositioning framework returns a list of subgroups as output. Each subgroup has a contrast score, the SPC_{Score}, representing the contrasts between a given subgroup and the entire disease population. For each subgroup, a set of drugs within each subgroup's network are ranked using $D_{Score,}$ where an increased score represents a drug more relevant to the subgroup. In each stage of the analysis, medical experts, the physician co-authors specializing in breast, colorectal, and lung cancers, were included in the decision process. In the evaluation for the candidate drugs, a physician-in-the-loop is required to evaluate both the effectiveness and side effects of top-ranked drugs. Physicians can assess the effectiveness of a drug based on the relationship between the drug's molecular profile as well as the gene patterns in the perturbed biological entities. Such explainable results, contributed from the motivation and design of the algorithm, are intuitive to clinicians when explaining why the drugs are recommended with underlying biological mechanisms to healthcare providers, as explainability is a critical limitation to the adoption of many current data mining methods. For the candidate drugs, the physician

can further evaluate pateint comorbidity, risk factors, and medical history to assess for interactions and potential side effects. This section has been published in the Journal of Biomedical Informatics [43].

2.4.2 DRUG EVALUATION USING COMPREHENSIVE DRUG SCORING

The drug scoring function was upgraded to include different factors that put the drug impacts within the subgroup network (SGNW) into perspective, as well as their impacts in general on human cells. The number of abnormally expressed genes in a subgroup, gene expression affected by each drug, and the importance of each gene within the SGNW are considered. For example, hub genes have more significance than genes with few connections. Gene weight is based on the percentage of entities of each biomedical type that connects to each gene. The mean of gene frequency (*MGF*) is calculated for each gene in the SGNW, as follows:

$$\mathrm{MGF}(Gene_i) = \frac{\sum_{k=1}^{e} \frac{nE_{ki}}{NE_k}}{e} \tag{9}$$

, where $Gene_i$ is a gene in SGNW for which the frequency is calculated, e is the total number of entity types in S, and $E_k \in S$, $S = [Gene, Pathway, \text{Biological process}, \text{Cellular component}, \text{Molecular function}]$, and nE_{ki} is the total number of entities of type E_k that are in direct interaction with $Gene_i$. NE_k is the total number of entities of type E_k that exist in the SGNW. The MGF is calculated for all the genes in the SGNW.

The algorithm calculates an initial weight for each drug, the accumulative gene frequency (*AGF*), which corresponds to the accumulative weight of all the genes connected to each drug in the SGNW.

$$AGF(Drug_j) = \sum_{i=1}^{m} MGF(Gene_i) \tag{10}$$

, where m is the total number of genes whose expression is altered by $Drug_j$ within a given subgroup. The AGF for $Drug_j$ is the overall summation of the MGF scores for each of these genes. This equation assigns a value that represents the importance of the drugs in terms of their ability to perturb gene expression. The weight represents the importance of a gene in the subgroup network. The AGF value characterizes a drug's importance as the average of the genes' importance connected to that drug.

The other significant factor to consider is the patterns of each subgroup. In the algorithm, genes are not treated as independent entities. In addition to taking the interaction of the subgroup genes with other genes and other biomedical entities, the gene patterns formed within each subgroup are considered. To address this, the percentage of patterns targeted by each drug, $PA(Drug_j)$, is calculated as follows:

$$PA(Drug_j) = \frac{NP_j}{NP} \tag{11}$$

, where NP_j is the total number of gene expression patterns targeted by NP_j in the subgroup of interest, and NP is the total number of gene expression patterns in this subgroup. Then, the overlap score measure (OSM) is calculated for each drug in the SGNW. The OSM of a drug is the summation of that drug's importance on two levels. One level is the connection between that drug and the genes in the subgroups, AGF. The other level is the importance of the drug based on the patterns targeted by that drug, PA:

$$OSM(Drug_j) = PA(Drug_j) + AGF(Drug_j). \tag{12}$$

In addition to the accumulative gene weights and the percentages of gene patterns impacted by each drug, the algorithm considers the ratio of genes whose expression is affected by each drug. This is accomplished by calculating the gene percentage targeted by each drug (GP) as follows:

$$GP(Drug_j) = \frac{NG_j}{NG} \quad (13)$$

, where NG_j is the total number of genes whose expression is perturbed by $Drug_j$ in the subgroup of interest, and *NG* is the total number of genes in this subgroup.

In this study, the drugs with high a OSM are essential to a given subgroup, and the prioritization of these drugs should be ensured. This is similar to the idea in information retrieval theory, where terms that appear in the majority of documents have less importance, like stop words such as "the", and "is", and so on. Drawing inspiration from idea of the inverse document frequency (IDF) to calculate the drug score [65], the inverse drug frequency (IDF) is calculated for each drug in each subgroup. This helps to increase the score of drugs that are unique to a subgroup and decrease the score of common drugs that could be related to cancer in general, but not to that subgroup of interest in particular. The biological reasoning behind using IDF as a scoring factor is to account for the amount of perturbation a drug could cause in the human cells. The drugs that most of its targets are part of the subgroup's gene patterns need to be preferred to prevent causing unnecessary perturbation in the human body. The IDF was calculated as:

$$IDF(Drug_j) = \log_{10}\left(\frac{Ns}{nS_j}\right) \quad (14)$$

, where Ns is the total number of genes in the RD-KB and nS_j is the number of genes targeted by $Drug_j$ in that subgroup network.

Finally, the impact of each drug is evaluated within the SGNW to rank the drugs based on different factors that determine each drug's effect. The final drug score ($D_{Score}(Drug_j)$) is:

$$D_{Score}(Drug_j) = OSM(Drug_j) * IDF(Drug_j) * GA(Drug_j). \quad (15)$$

Each drug within the SGNW is assigned a D_{Score} for ranking purposes. The higher the score, the more potential the drug has as a treatment for a given subgroup. This method addresses the importance of considering different factors in ranking drugs. The ranking is not only based on the genes connected to the drugs and their significance in the network, but also on the gene expression perturbation a drug can cause in human cells and the number of gene patterns affected by the drug within the subgroup's network. This section has been published in the Cancers journal [66]

Chapter 3
Results

3.1 INTRODUCTION

The results portion reviews the outcomes of three experiments which implement the subgroup prioritization and drug repositioning framework. Section 3.2 is the results of implementing the patients stratification and aggregated drug scoring on colorectal cancer data. Section 3.3 is the results of implementing the patients stratification and comprehensive drug scoring on breast cancer data. Section 3.4 is the results of pan-cancer analysis after implementing the patients stratification and comprehensive drug scoring on a set of 11 cancer types that have a survival rate of less than 75%. These cancer types are brain, colon, esophagus, kidney, liver, lung, ovary, pancreas, rectum, stomach, and triple negative breast cancer. The TCGA was the source for the genotypic and phenotypic data for the diseases in these three experiments.

3.2 COLORECTAL CANCER ANALYSIS RESULTS AND DISCUSSION

Differentially expressed genes and the phenotypic data were used with the heterogeneous biomedical knowledge base for subgroup discovery and drug repositioning using CRC patients from the TCGA. This section has been published in the Journal of Biomedical Informatics [43]. In this section, the results of the analysis are explained using CRC as a case study. Section 3.2.1 describes the subpopulation discovery results. Section 3.2.2 concerns drug repositioning and prioritization for each subgroup within this CRC population. Section 3.2.3 discusses the relations between drugs and genes in the subgroup networks. Section 3.2.4 presents the analysis of the subgroups' top-ranked drug candidates. Section 3.2.5 is the discussion of the finding in the colorectal cancer SPDR analysis.

3.2.1 SUBPOPULATION RESULTS

After performing exploratory data mining, the resulting subgroups were filtered based on the SPC_{Score}. Subgroups with $SPC_{Score} > 0$ are considered to be significant CRC subgroups. The total number of subgroups that met this condition was 130 (see Supplement 2 in [67]). The SPC_{Score} ranged between 2.5 and 53 (mean $SPC_{Score} = 17.44 \pm 8.60$; see Table 1 and Supplement 2 in [67]). These subgroups were then categorized based on clinically relevant features in consultation with clinicians experienced in treating CRC patients. Among 130 subgroups, the focus was directed to a set of subgroups with population variables containing microsatellite status (MS) which has three possible values/categories, namely, microsatellite instability-high (MSI-H), microsatellite instability-low (MSI-L), and microsatellite stability (MSS). Microsatellite status is a critical clinical feature of CRC, as studies have demonstrated important molecular differences impacting treatment response [68]. Additionally, patients with MSI-H tumors are the only patients with CRC that have demonstrated significant responses to immune checkpoint blockade [69]. Using this sub-categorization, 25 subgroups were found with MS status as part of the population variables listed in Table 1.

Subgroup category	Number of subgroups	Number of population variables		SPC_{Score}		Number of patients per subgroup		Genes per group		Number of drugs per group	
		Min	Max	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MSI-H	11	1	3	37.85	10.39	64	12	642.55	518.88	845.27	267.39
MSI-L	2	2	3	4.21	2.47	54	4	1045.5	183.14	1135.5	36.06
MSS	12	1	4	15.22	14.96	130	100	1234	701.28	1125.40	213.80

Table 1: Categories of MSI test subgroup s. MSI-H = microsatellite instability-high; MSI-L = microsatellite instability-low; MSS = microsatellite stable

In addition to MS status, there are critical differences in treatment outcomes based on specific clinicopathologic factors such as gender, anatomic location, and lymphatic invasion [70-73]. These factors were also found to be critical features using the exploratory data mining algorithm when combined with MS status. Due to clinical significance, the focus was on specific subgroups with these critical and clinically relevant features for in-depth study (see Table 1, Table S2 in Supplement 1, and Supplement 3 in [67]). To pictorially present subgroups from the three categories, Figure 4 shows subgroups matched with relevant drugs. For example, within the subgroups that have MSI-H as a phenotypic variable, right-sided colon cancer (P2R) and no lymphatic invasion (P6N) are features in three different subgroups. The first subgroup has MSI-H and P2R as phenotypic features and the suggested drugs for this subgroup are Cerulenin, Crizotinib, and Afatinib. The second subgroup has MSI-H and P6N as phenotypic features and the suggested drugs for this subgroup are Idarubicin, Dactinomycin, and Doxorubicin. The third subgroup has MSI-H, P2R and P6N as phenotypic features and the suggested drugs for this subgroup are Menadione, Dasatinib, and Vinblastine.

3.2.2 DRUG REPOSITIONING RESULTS

Unique, differentially expressed genes for each subgroup were used to query the drug knowledge base (see Table 1). For each differentially expressed gene (DEG), all the biomedical entities were retrieved to create a network that represents the given subgroup (see Table S3 in Supplement 1, and Supplement 3 in [67]).

1- Subgroup1 (SG1): The patients in this subgroup have MSI-H as their MSI test result and have no lymphatic invasion. This subgroup contains about 10% of patients. The uniquely differentially expressed genes are RAB37, GCK, and NOP56. The top

three predicted candidate drugs for this subgroup are Idarubicin, Dactinomycin, and Doxorubicin.

2- Subgroup2 (SG2): The patients in this subgroup have MSI-H as their MSI test result, and they have right-sided colon as their anatomic neoplasm subdivision. This subgroup contains about 14% of the sample. The uniquely differentially expressed genes are PLXNA1, FDXR, AGRN, and MYO7A. The top three predicted candidate drugs for this subgroup are Cerulenin, Crizotinib, and Afatinib.

3- Subgroup3 (SG3): The patients in this subgroup have MSI-H as their MSI test result. These patients have right-sided colon as their anatomic neoplasm subdivision with no lymphatic invasion. This subgroup contains about 10% of the sample. The uniquely differentially expressed genes are TCEA2, SNRPN, SPG20, LOC157381, YPEL4, MUM1L1, FBXO17, MYLK3, NHLRC1, ELMOD1, COL25A1, CBLN4, LOC339535, SOX30, and KCNJ3. The top three predicted candidate drugs for this subgroup are Menadione, Dasatinib, and Vinblastine.

4- Subgroup4 (SG4): The patients in this subgroup are female with MSS as their MSI test result. This subgroup contains about 29% of the sample. The uniquely differentially expressed genes are PRKY, GYG2, XIST, and COX7B2. The top three predicted candidate drugs for this subgroup are Varenicline, Digitoxin, and Gefitinib.

5- Subgroup5 (SG5): The patients in this subgroup are female with MSS as their MSI test result, and they have lymphatic invasion. This subgroup contains about 12% of the sample. The uniquely differentially expressed genes are CPT2, FNIP2, PANK3, SIPA1L2, PPARGC1B, GCET2, RAB27B, ALDH1A1, EPHA4, SLITRK6,

UGT2B7, GAS2L2, KLK3, DEFA5, C1orf112, RPL23AP7, MFRP, NOS3, ARSE, TBX6, TNFRSF4, FCRLB, SUSD3, MYL4, AQP7P1, SNHG9, MMP17, MPO, C10orf82, ART5, NKAIN4, PCDHA4, UPB1, PRINS, PLSCR2, MLANA, PKHD1, C14orf53, ZPBP2, HBG1, PCDHA12, KIAA0087, LOC100133469, and PCDHA11. The top three predicted candidate drugs for this subgroup are Crizotinib, Cerulenin, and Dabrafenib.

6- Subgroup6 (SG6): The patients in this subgroup have **MSS** as their MSI test result, and they have **a history of colon polyps**. This subgroup contains about 20% of the sample. The uniquely differentially expressed genes are CACNG8, OR10H1, LOC440173, DUXA, KRTAP10-6, KCNA10, FGF21, SSX7, CACNG1, KRTAP5-7, HIST1H2BF, GPR144, GOLGA9P, INGX, DSPP, P2RX3, EPX, RNF222, KRTAP10-12, LOC100128675, LCNL1, FAM75A3, KRTAP10-2, NPBWR1, GPR152, FAM75A6, C14orf166B, TAS1R2, SLC22A8, RGS7, PTX4, FLJ42393, and RBMXL3. The top three predicted candidate drugs for this subgroup are Niclosamide, Perhexiline, and Digoxin.

7- Subgroup7 (SG7): The patients in this subgroup have MSS as their MSI test result, have a history of colon polyps, and they do not have venous invasion. This subgroup contains about 9% of the sample. The uniquely differentially expressed genes are NDRG1, NLGN3, TGFBR3, GABRB3, RPS5, ROMO1, SNAR-C3, and C9. The top three predicted candidate drugs for this subgroup are Menadione, Varenicline, and Crizotinib.

In this analysis, the focus was on the top 3 drugs for each of the seven subgroups. The total number of drugs was 16. Twelve of these 16 drugs are cancer-related therapies. Eight of these twelve are also associated with CRC treatment. These drugs are Dasatinib, Crizotinib, Niclosamide, Dabrafenib, Gefitinib, Afatinib, Doxorubicin, and Menadione. These drugs were used or suggested as either mono or combined therapy for CRC patients with different genetic features.

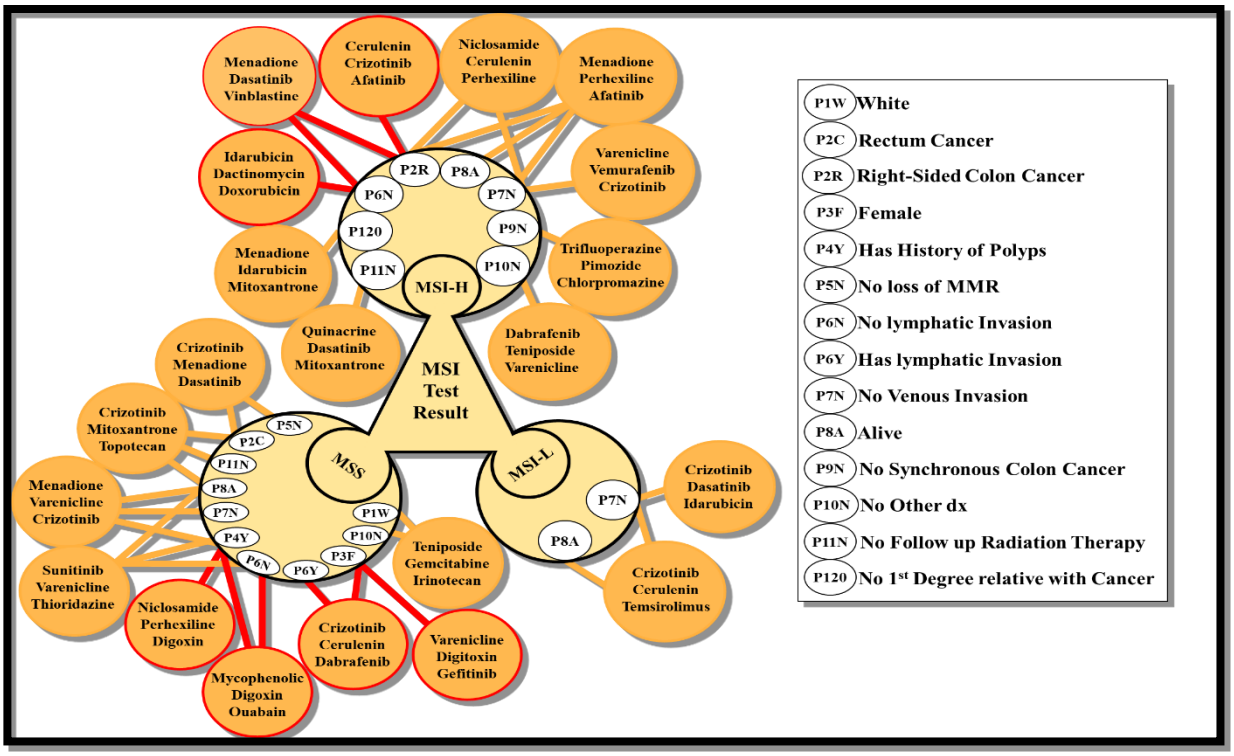

Figure 4: MSI test result-related subgroups with the top three recommended drugs for each subgroup. The white circles represent the phenotypic properties, and the golden circles represent the recommended drugs for a subgroup that have the phenotypic properties that are connected to the given drugs' circle in addition to the MSI test result. The phenotypic variables connected to drugs circled with a red border are the subgroups that are highlighted. MSI-H = microsatellite instability-high; MSI-L = microsatellite instability-low; MSS = microsatellite stable; MMR = mismatch repair; dx = diagnosis

3.2.3 DRUG-DIFFERENTIALLY EXPRESSED GENE SUBGROUP NETWORKS

Different biomedical entities in a subgroup network have different roles in deciding the drug and drug ranking within the network. The genes and gene interactions aid in

determining the extracted drugs, while other biomedical entities factor in weighting the genes and ranking the drugs. The analysis for the genes that caused drugs to be suggested for a subgroup showed the importance of pathway and gene interactions included in the study. DEGs in a subgroup interact with different drug targets and genes that are affected by the treatment to create the molecular profile of a drug in the given subgroup. For example, Menadione was suggested in subgroup3 and subgroup7 because it affects genes that have direct interactions with the DEGs in these subgroups. Menadione is a form of vitamin K that plays a critical role in blood clotting and bone health. Based on *in vitro* cell line investigations, Menadione was found to have anti-cancer effects, including in CRC [74, 75]. In subgroup3, Menadione downregulates 19 genes and upregulates 17 of the genes that interact with six DEGs unique to subgroup3 (Figure 5). In subgroup7, Menadione downregulates 36 genes and upregulates 25 genes that have direct interaction with six of the DEGs for that subgroup (Figure 5).

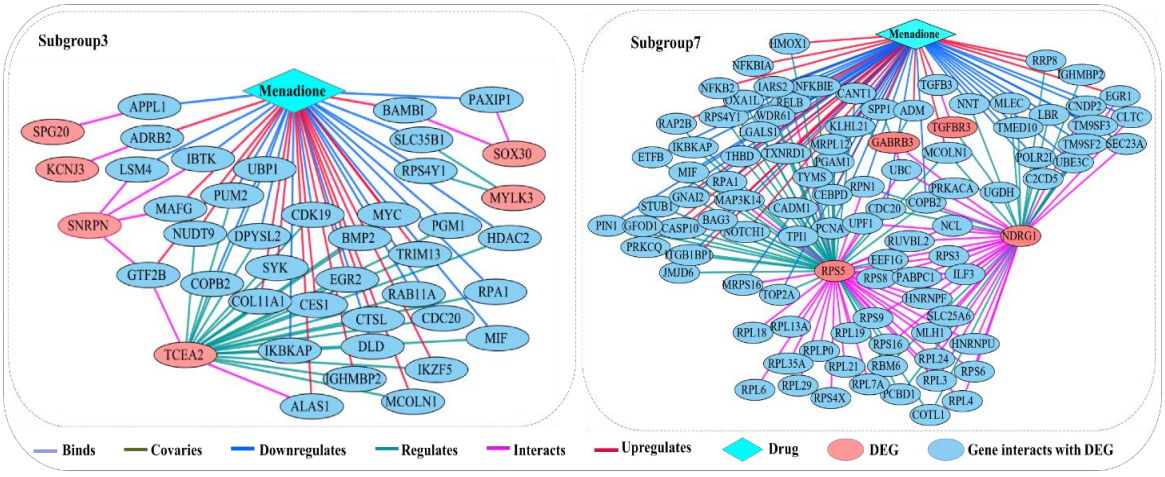

Figure 5: Menadione gene interactions in two different subgroups

For the top three drugs in these seven subgroups, Crizotinib was the one most frequently suggested (3/7). Crizotinib was recommended for subgroup2, subgroup5, and subgroup7 (Figure 6). As expressed in DrugBank, Crizotinib is “an inhibitor of receptor

tyrosine kinase for the treatment of non-small cell lung cancer (NSCLC).” Some studies have also investigated its potential in CRC patients. When used for combination therapy in MSI-H, BRCA2 deficient patients with c-MET overexpression, Crizotinib was found to increase apoptosis and tumor cell death[76]. In subgroup2, where Crizotinib was recommended for patients who have MSI-H right-sided CRC, MET gene expression was upregulated, and mutated in 28% of patients. For CRC tumors where SOX13 mediates cells migration, invasion, and metastasis, it has been found that inhibiting c-MET using Crizotinib prevents CRC metastasis by blocking HGF/STAT3/SOX13/c-MET axis [77] and SOX13 is mutated in 19% of patients in this subgroup. Crizotinib was found to have a role in overcoming the resistance to some drugs like Cetuximab. In some CRC cell lines, the resistance was developed to Cetuximab as a result of activating Tyrosine Kinases (RTKs) like MET and RON, or the resistance that resulted from adding HGF and NRG. Adding Crizotinib to these cell lines blocked resistance to Cetuximab [78]. Also, crizotinib has also been shown to overcome resistance to Cetuximab and improve the chemoradiation outcome in CRC cell lines that carry mutant KRAS [79]. In subgroup5, where the patients are females that are MSS and have lymphatic invasion, MET is mutated in 15% of patients and gene expression is upregulated. BRCA2 is also upregulated and mutated in 9% of these patients. In subgroup7 where the patients have MSS and a history of colon polyps with no lymphatic invasion, MET is upregulated and mutated in 23% of patients with STAT3 mutated in 9% of patients. All these data support the recommendation of Crizotinib for these patients and demonstrate the power of this explainable data mining application. As was observed with Menadione and Crizotinib, a drug can be recommended for more than one subgroup though DEGs are different, and the genes set by drug interactions are

different in different subgroups. This shows that the proposed method for hypothesis generation can be used to recommend drugs based on the molecular profile for patients within each subgroup by matching the gene signature of the drug and the gene signatures of patients in identified subgroups. Also, when comparing drugs within the same subgroup of patients taking Menadione and Crizotinib in subgroup7 as an example, a set of genes can be the same in both drugs' networks. However, to decide which drugs should be used for which patients within that subgroup, the unique gene signature must be found in each drug's network. In subgroup7, about 78% of the genes in the Crizotinib network are unique to Crizotinib, while in Menadione's network, about 84% of the genes are unique to Menadione's network.

The validation of each drug recommended for each subgroup is explained in Table S3 of Supplement 1 in [67], where each row in the table represents the matching between the drug profile and the genotypic features of the patients in a given subgroup from the literature.

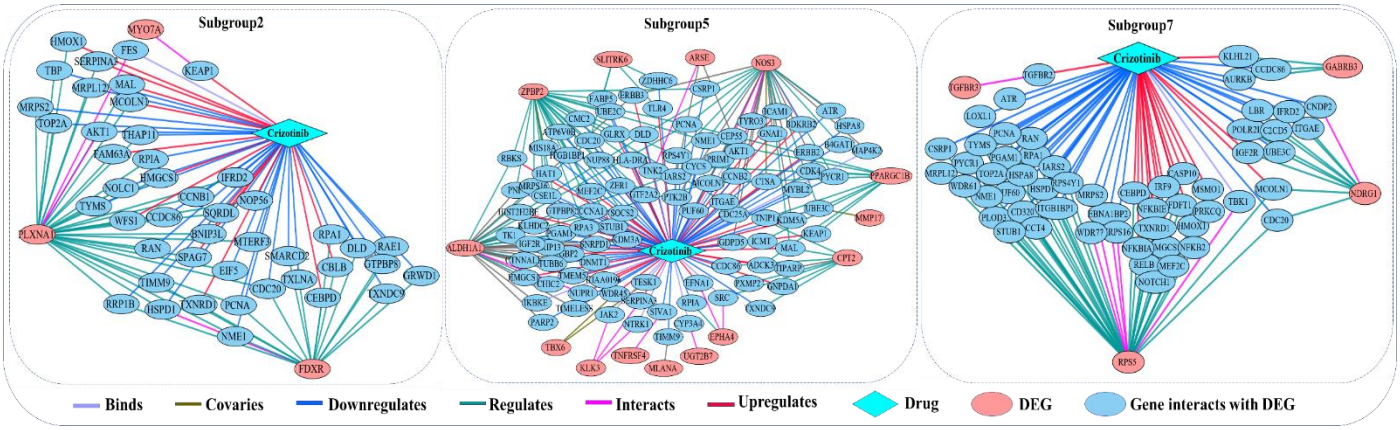

Figure 6: Crizotinib gene interactions in three different subgroups

3.2.4 ANALYZING THE SUBGROUPS' TOP-RANKED DRUG CANDIDATES

For each candidate subgroup, the top-ranked drugs based on D_{Score} were analyzed computationally and biomedically. First, randomized analysis was performed to ensure these drugs were not suggested for each subgroup by chance, and that there is a biomedical meaning that could be explored (Section 3.2.4.1). Then, these drugs were analyzed in the biomedical context using drug class enrichment analysis (Section 3.2.4.2) and pathway enrichment analysis (Section 3.2.4.3) to find the biomedical effect of these drugs.

3.2.4.1 RANDOMIZED ANALYSIS

For validation, a randomized analysis was conducted within each subgroup. For each subgroup's network, the nodes were preserved and the edges were shuffled to create a random network. For each subgroup, ten random networks were created. Then, the drug repositioning method was performed on each network. The results showed that the top drugs generated based on the randomized networks were different from the top 15 drugs generated from the original subgroup networks. Indeed, there were only two candidate drugs that were ranked 18th and 34th from top drug list in the randomized network. The

Subgroups	DR Drugs	Randomized Drugs
MSI-H & no lymphatic invasion	Idarubicin	Tetrahydrobiopterin
	Dactinomycin	Diethylstilbestrol
	Doxorubicin	Fluphenazine
MSI-H & right-sided colon	Cerulenin	Streptozocin
	Crizotinib	Tetrahydrobiopterin
	Afatinib	Metaxalone
MSI-H & right-sided colon & no lymphatic invasion	Menadione	Testosterone
	Dasatinib	Pseudoephedrine
	Vinblastine	Paliperidone
MSS & female	Varenicline	Terazosin
	Digitoxin	Isoflurane
	Gefitinib	Fospropofol
MSS & female & lymphatic invasion	Crizotinib	Imatinib
	Cerulenin	Simvastatin
	Dabrafenib	Bortezomib
MSS & a history of colon polyp	Niclosamide	Trilostane
	Perhexiline	Doxylamine
	Digoxin	Flucloxacillin
MSS & a history of colon polyp & No venous invasion	Menadione	Diclofenac
	Varenicline	Digoxin
	Crizotinib	Progesterone

Table 2: The results of the randomized analyses listing the top three drugs of our DR method and the top three drugs of the randomized networks for each of the selected subgroups.

remaining drugs were ranked greater than 40. Table 2 shows the top three drugs from both methods in detail. The ranking of the drugs in the random networks was based on the average rank for each drug in a given subgroup's networks.

3.2.4.2 DRUGS' CLASSES ENRICHMENT ANALYSIS

The class of each drug among the top 10 repositioned candidates was examined to evaluate the candidate drugs recommended for the seven subgroups. After obtaining the top 10 drugs for each subgroup and removing duplicates, 37 unique drugs were found. Then, these drugs were mapped to the knowledgebase to link the pharmacological class for each. There were 31 different classes enriched in these drugs. The top five pharamcological classes are shown in Figure 7.

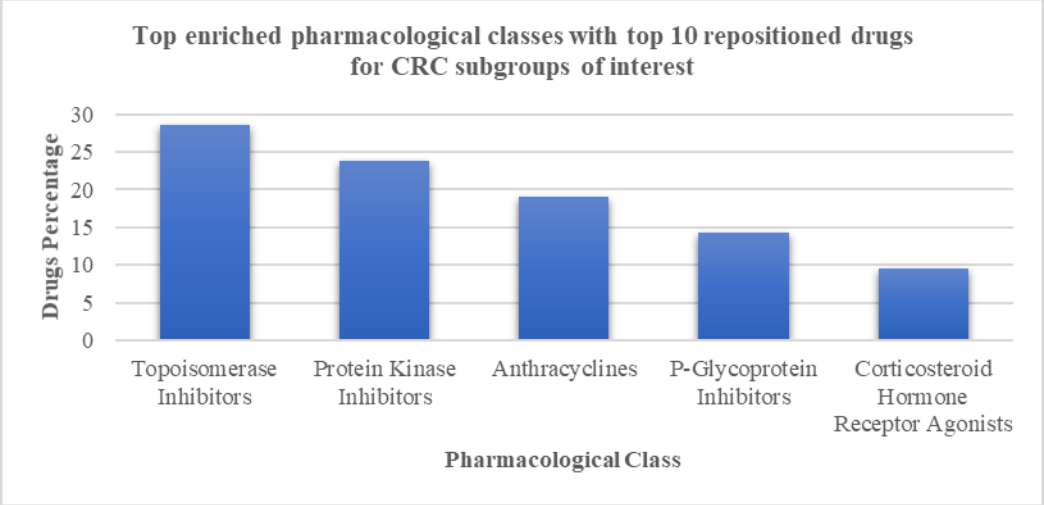

Figure 7: Top pharmacological classes that were highly enriched with the drugs repositioning candidates for the seven subgroups of interest.

The remaining classes were enriched in less than 5% of the drugs. The result shows that Topoisomerase Inhibitors, Protein Kinase Inhibitors, Anthracyclines, P-Glycoprotein Inhibitors, and Corticosteroid Hormone Receptor Agonists were the pharamcological classes that were highly enriched with the top 10 drugs for the seven subgroups of interest.

Topoisomerase Inhibitors, such as Irinotecan, a drug commonly used in the treatment of CRC, are considered some of the most effective apoptosis inducers [80]. This is due to their ability to target Topoisomerase enzymes, which have a significant role in DNA replication [81]. The sensitivity to these inhibitors was also found to be increased in colorectal cancer cell lines defective in DNA MMR, a critical patient group [82].

The second class is protein kinase inhibitors, which have been developed to block pathways related to tumor growth and progression. Studies have shown that these inhibitors have potential to be used in the treatment of metastatic colorectal cancer [83]. Anthracyclines have been shown to be effective for the treatment of breast cancer with TOP2A mutations, and patients with metastatic CRC were found to have a higher rate of mutation in this gene than in breast cancer, leading to a phase II trial in CRC [84, 85]. Different P-Glycoprotein Inhibitors have been developed or recommended as repurposed drugs to treat cancer [86]. This shows that the majority of our drugs belong to cancer-related classes and have the potential to be repositioned for colorectal cancer.

3.2.4.3 PATHWAY'S ENRICHMENT ANALYSIS

To understand the common underlying mechanism of action for the top drugs from all the subgroups of interest, signaling pathways that are highly enriched with these drugs' gene targets were analyzed. For the top 10 drugs of each subgroup, a gene set enrichment analysis was performed to find the highly enriched pathways for each drug's targets [87, 88]. For the 37 drugs, the top 10 pathways for each drug were examined. A summary of findings is shown in Figure 8, where the x-axis represents the top pathways that are highly enriched in the repositioned drug candidates, and the y-axis represents the percentage of drugs in which a given pathway was within the top 10 pathways targeted by a drug. The

pathway targeted by about 80% of these drugs was the cell cycle pathway; then, the p53 pathways followed chronic myeloid leukemia. Different drugs were developed and have been shown to produce anticancer effects affecting signaling pathways, including the cell cycle.

Drugs like proteasome inhibitors have shown effectiveness on human CRC cells [89].

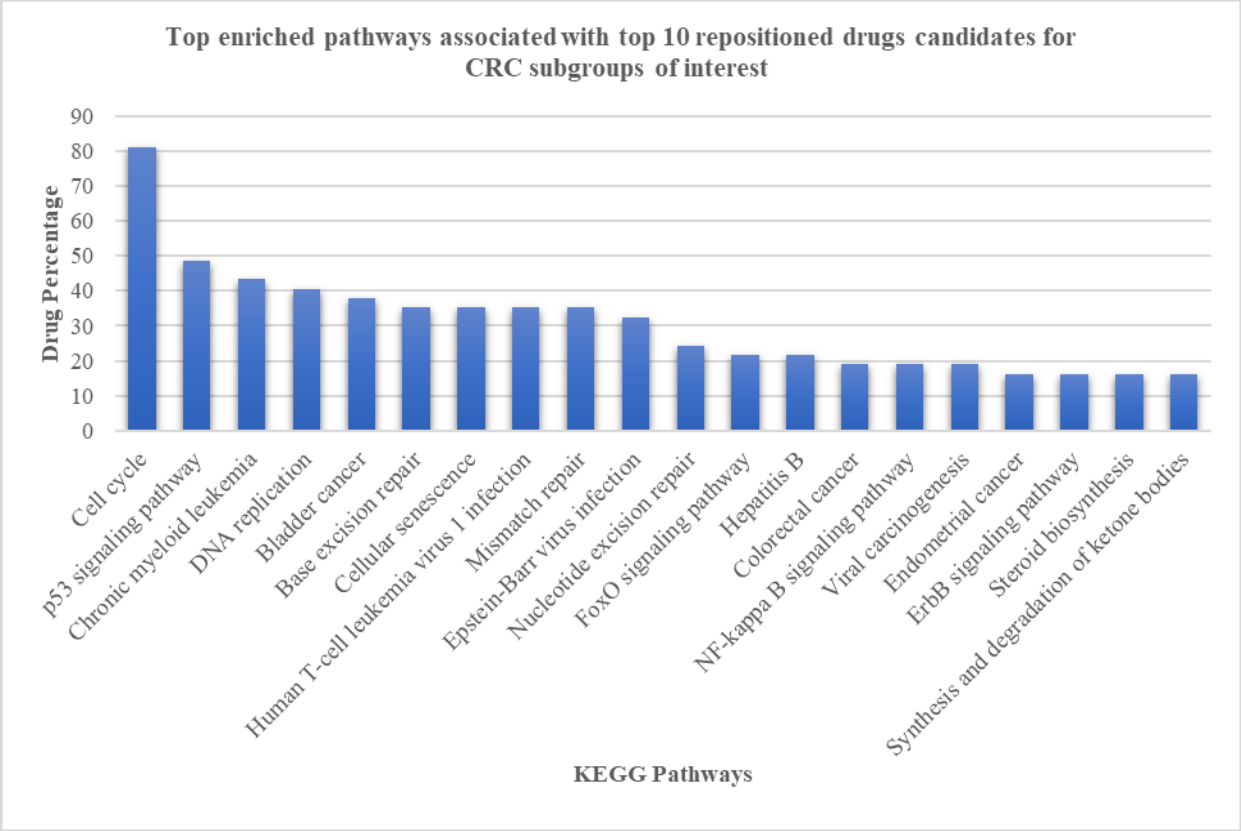

Figure 8: Top pathways that were targeted by repositioned drug candidates for the seven subgroups

Another study revealed that affecting cell cycle components inhibits colorectal cancer cell proliferation [90]. The second pathway is the p53 pathway. P53 is a key tumor suppressor gene that is mutated or lost in different cancer types including CRC. Regulating the p53 pathway impacts apoptosis [91]. Also, some studies have suggested that restoring the p53 pathway may enable selective cell death in cancer, including CRC, in which the cancer cells can be targeted without affecting the normal cells [92]. The third pathway was chronic

myeloid leukemia. A study found that a c-kit tyrosine kinase inhibitor with a significant effect on chronic myeloid leukemia could be repurposed for colorectal cancer patients expressing the c-kit proto-oncogene [93]. In this study, and after testing the inhibitor in human colorectal tumor cells in vivo and in vitro, the inhibitor has the potential to prevent colon cancer and to treat advanced CRC related to liver metastases. Additionally, the mismatch repair pathway, a common dysregulated pathway in CRC, and the colorectal cancer pathway were highly targeted pathways by these drugs. These results show that the recommended drugs for the subgroups of interest are highly enriched for targeting cancer-related pathways.

3.2.5 DISCUSSION

In this study, the framework for disease population stratification and drug repositioning was implemented on CRC data. Its explainable results can be used to generate hypotheses for future clinical trials in which researchers and physicians can tailor treatments to subgroups of patients through the automatic determination of inclusion and exclusion criteria and the unique molecular profile of each drug within a subgroup. The subgroups were selected based on phenotypic features and genotypic patterns. The differentiation between a subgroup and its outer population is based on the contrast using these patterns and network analysis. Part of the explainability of the findings is that the SPC_{Score} and D_{Score} can be traced back to the basic biomedical components supporting the selection of the repurposed drugs. As different research has shown the possibility of non-cancer related drugs, such as those for diabetes, to be repositioned for cancer treatment [94-98], the results also recommended drugs that are non-cancer related but may potentially be used for CRC patients. This is in addition to the majority of identified drugs that were

originally designed for cancer-related therapy but mostly not in current use in CRC. For the drugs recommended to be repositioned, but presently not directly associated with CRC treatment, the abnormal genes of the given subgroups were found to directly or indirectly interact with the targets of these drugs, which had the highest aggregated scores in the subgroup's network. Additionally, wet lab-based *in vitro* experiments demonstrated that some of these drugs have a potential role in CRC treatment as a single or combinatorial therapy that may overcome resistance to other CRC drugs [74, 79, 99-104]. In this analysis of the top drugs for the subgroups of interest, enrichment analysis for these drugs was done based on their pharmacological class and pathways that are highly enriched as targets for the drugs. Topoisomerase inhibitors were the pharmacological class that had the highest number of drugs. Topoisomerase inhibitors affect the cell cycle and result in cell death. In addition to drugs that are already approved as CRC treatment in this class, others have demonstrated activity in metastatic colorectal cancer therapy [105]. Also, it can be employed as a combined therapy to increase programmed cell death in CRC [106]. Regarding the analysis for the pathways, the top pathways targeted by most top drugs were cancer-related pathways. While published literature was used to validate the findings in addition to the randomized analysis that showed the recommended drugs were not selected by chance, further wet lab experiments to validate these findings are required prior to initiating clinical trials using these repurposed therapies.

3.3 BREAST CANCER ANALYSIS RESULTS

Breast cancer (BC) is the leading cause of death among female patients with cancer [107, 108]. BC is a highly heterogeneous disease. BC has intertumoral heterogeneity, where the tumor is heterogeneous among the BC patients, and intratumor heterogeneity, where the

tumor is heterogeneous within the tumor of an individual patient [109]. A consequence of heterogeneity is that patients with breast cancer can have different reactions to the same drug. BC is a clear example of the need to have personalized medicine that is patient-centric rather than disease-centric. Implementing precision medicine in a healthcare system requires developing patient stratification methods to find homogeneous subgroups of patients to tailor drugs to these subgroups. Developing de novo drugs is time-consuming, and is expensive process [1, 2]. To overcome the problems associated with developing drugs for a subpopulation of a disease, drug repositioning (DR), the redirection of already approved drugs to be used for additional diseases by finding new indications, emerges as an alternative to support precision medicine implementation [9].

Different molecular features have been used to stratify patients and reposition drugs. These molecular features range from using a single gene to using heterogeneous molecular data types; for example, using P53 to stratify patients based on their P53 status and using the RACK1 status to determine personalized treatment [110]. Moreover, heterogeneous molecular data was used for cancer subtyping and the repositioning of drugs after ranking genes based on the Gene Ontology (GO) pathways analysis [111]. The transcriptional response to drugs was used to infer signaling interactions and to reposition drugs [112]. Drug repositioning was also accomplished by targeting pathways that play a role in cancer proliferation and progression [113]. The oncogenic PI3K-dependent inhibitor was recommended as a cancer therapy [114]. Drugs were recommended for repositioning based on the similarity between a particular cancer type therapy and a given drug using a drug-pathway network [27]. The association between gene mutation and expression was also used to guide the drug discovery process [115]. A network analysis was used widely to reposition

drugs for various types of cancer. Studies have introduced many heterogeneous network models to reposition drugs, including the drugs-genes-diseases network [116, 117], and drug-targets-pathways-genes-diseases, where the association between drugs and diseases was found based on multiple targeted genes and pathways between drugs and diseases [118]. The drug-disease association was also inferred based on a set of heterogeneous networks including drug-gene, disease-gene, protein-protein, and genes co-expression networks to find drug-disease associations and to reposition drugs[119].

Several methods have been developed to understand the underlying mechanism of BC and the recommend drugs. A gene expression analysis was used to identify pathways involved in BC invasion and potential drug targets [120]. In the context of DR, drugs have been recommended for their antiproliferation effect on BC patients [121]. Additionally, drugs that share side effects with BC therapies have been recommended to be repositioned for BC [122]. A pathway analysis was used with single-cell data to recommend drugs for BC [123]. A network analysis also was used to reposition drugs for breast cancer. A tissue-specific protein-protein interaction (PPI) network was used to evaluate and reposition drugs over the drug-miRNA-diseases network [124]. A network propagation analysis was used to reposition drugs based on a drug-pathway network, where pathways for each drug were identified by an enrichment analysis of the genes regulated by each drug using the Connectivity Map (CMAP) for the drug's phenotypic profile [125]. Moreover, computational methods with diverse molecular data types were developed to identify therapeutic targets for BC [126]. These methods considered breast cancer in general without considering its subtypes and did not address the heterogeneity of BC. Other methods have been developed to stratify BC patients. This stratification was done based on cell receptor

statuses; for example, an estrogen receptor positive (ER+) status and an estrogen receptor negative (ER-) status [127].

Based on the gene expression profile, four molecular subtypes have been identified for breast cancer which are luminal A, luminal B, human epidermal growth factor receptor 2 (HER2)-enriched, and basal-like (triple negative) [128]. Computational methods have been developed to find therapeutic biomarkers and to reposition drugs for breast cancer subtypes. The correlation between DNA copy number alterations and gene expression was used to find biomarkers for each subtype [129]. Electronic health record (EHR) and differentially expressed genes (DEGs) data were used to reposition drugs for the four BC molecular subtypes by finding drug pairs using drug-protein relations [130]. mRNA was used to reposition drugs for these subtypes using gene co-expression [131]. Moreover, drugs were predicted based on the suggested miRNA biomarkers for each subtype [132]. Drugs were recommended based on the number of hub genes each drug targeted within the miRNA-protein-drug network [133]. An integrated network of relations between lncRNA, miRNA, and mRNA was used to recommend drugs that reversed the lncRNA expression of each BC molecular subtype [134]. A pathway analysis was also used to reposition drugs for BC subtypes [135]. Biomarker predictions using cell line data were used to stratify BC patients and suggested drugs for each subgroup [136]. These methods have demonstrated the importance of stratifying BC patients into homogeneous subgroups and repositioning drugs into these subgroups. Still, they have considered only the genotypic characteristics of BC patients without considering the significance of phenotypic features. Other methods have taken phenotypic data into consideration, but the majority of these methods include the phenotypic features as the post-analysis of the stratification process, where clustering

methods have been used to cluster the genotypic profiles of patients and then map the phenotypic traits to each cluster in order to assign patients to each subgroup [25].

This study implements our data-driven approach to stratify BC patients based on genotypic and phenotypic data. The genotypic data is mapped onto a heterogeneous drug knowledge base to find druggable homogeneous BC subgroups and to reposition drugs for each subgroup [67]. This study focuses on triple negative breast cancer (TNBC), as defined by the lack of Estrogen Receptors (ERs) and progesterone receptors (PRs) and by a HER2-negative status, due to its increasingly recognized heterogeneity not only on the molecular level, but also on the pathologic and clinical levels [137]. The lack of targets like ER, PR, and HER2 in TNBC implies that chemotherapy remains the only treatment of choice for patients with TNBC, which unfortunately fails to achieve prolonged remission in most cases. Indeed, on recurrence, patients with TNBC have worse survival outcomes than patients with ER-positive and/or HER2-positive BC subtypes [138, 139]. Other studies have demonstrated the importance of finding homogeneous subpopulations within the TNBC population, but these studies mainly focused on a set of genes or single nucleotide polymorphisms (SNPs) to identify drug targets for TNBC [140, 141]. Other methods had a broader range of data, such as gene expression data, to identify key genes that could become drug targets [142]. As these methods represent a step toward understanding TNBC progression and improving patient survival, targeting a heterogeneous disease requires considering a wide variety of genotypic and phenotypic factors. Therefore, the developed data-driven approach was implemented in order to dissect TNBC heterogeneity as much as possible and to discover druggable molecular targets that may enhance the chemotherapy

benefit by blocking chemoresistance pathways and, thus, cancer recurrence in TNBC patients. This section has been published in the Cancers journal [66]

3.3.1 DATA DESCRIPTION AND PROCESSING

Data from 980 breast cancer patients was obtained from The Cancer Genome Atlas (TCGA). The dataset consisted of phenotypic data and genotypic data. For the phenotypic variables, we used 19 clinical variables (Figure 9- Module 1 and Table 3). The continuous phenotypic variables were converted into categorical variables. The categorization of the clinical variables was done based on medical guidelines [143]. The genotypic variables were the RNA-seq data of these patients.

A differential analysis was done between 94 normal and 980 tumor samples. The normal and tumor data were downloaded from TCGA. Based on the differential analysis done using EdgeR, 1531 most differentially expressed genes with a p-value less than 0.05, a log2FC > +2 for upregulated genes, and a log2FC < -2 for the down-regulated genes were selected. Then, the genotypic data were normalized using log2 and categorized based on the z-score value for each gene in each patient. These genes were categorized into upregulated (z-score > 1), downregulated (z-score < 1), and normal (z-score between 1 and -1) genes. The drug repositioning knowledge base (DR-KB) was represented as a neo4j graph with a gene-centric schema, including relations between genes, pathways, molecular function, cellular components, biological processes, disease, and drugs (Figure 9- Module 2). [24].

The method consisted of two parts. The first part was the identification of the homogeneous sub-groups within a heterogeneous disease population. To find the homogeneity between patients in a subgroup, genotypic pattern mining was performed for patients sharing phenotypic features. The relations between gene patterns and biomedical

entities were extracted from the DR-KB after mapping the gene patterns to the DR-KB graph. The second part of the method was to find suitable drugs for each subgroup by developing and applying a drug repositioning algorithm. This algorithm used graph analysis to analyze each subgroup's network. The drugs in each network were then ranked according to various factors (as discussed in Section 2.4.2) to consider a wide range of drug properties and impacts and to find a suitable regimen for each subgroup based on their genotypic features.

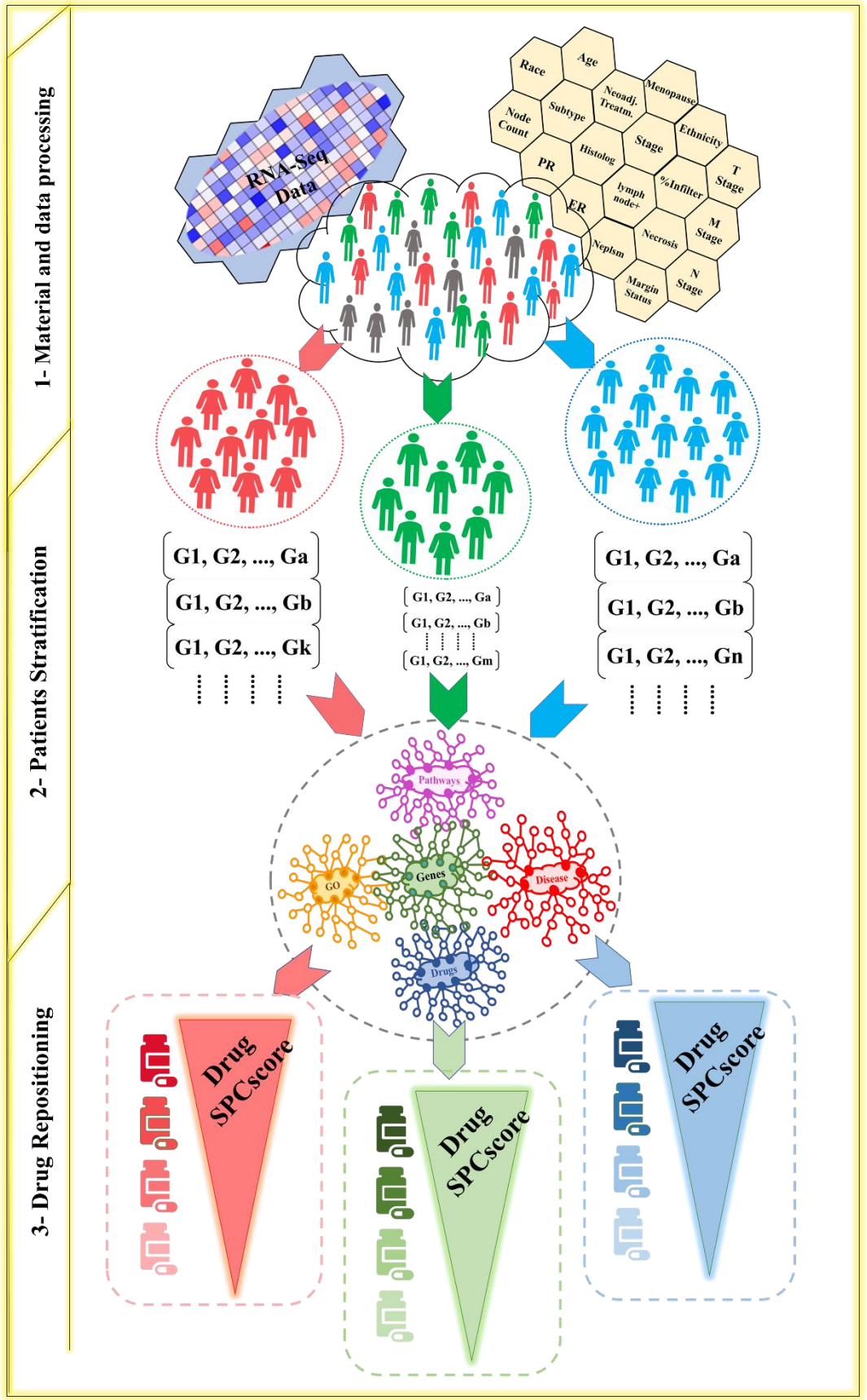

Figure 9: Flowchart of the data-driven drug repositioning process using phenotypic and genotypic breast cancer data of TCGA, patient stratification methods, and a knowledge base for recommendation of drug repositioning.

	Variable	Categories	Count
1	**Age**	<50	265
		>=60	459
		50s	256
2	**ER Level Cell Percentage Category**	0-49%	104
		50-99%	302
		None	574
3	**Histological Type**	IDC	702
		ILC	182
		infiltrating carcinoma nos	1
		MED	6
		metaplastic carcinoma	10
		MIXED	25
		MUC	15
		OTHER	39
4	**History of Neoadjuvant Treatment**	NA	1
		no	972
		yes	7
5	**Lymph Node Examined Count**	<=3	252
		>=10	429
		4-9	195
		None	104
6	**Margin Status**	close	28
		NA	64
		negative	888
7	**Menopause Status**	indeterminate	26
		NA	72
		peri	36
		post	642
		pre	204

8	**Neoplasm Subdivision**	left	508
		right	472
9	**Number of Lymph Nodes Positive**	<=3	691
		>=10	54
		4--9	91
		none	144
10	**Patient Ethnicity**	hispanic or latino	34
		NA	155
		not hispanic or latino	791
11	**Patient Race**	american indian or alaska native	1
		asian	59
		black or african american	156
		NA	81
		white	683
12	**Percent Lymphocyte Infiltration**	>40%	16
		0-9%	297
		10-39%	71
		none	596
13	**PR Level Cell Percent Category**	0-49%	193
		50-99%	182
		None	605
14	**Stage**	advanced	152
		nx	6
		primary	822
15	**Stage m**	m0	828
		mx	152
16	**Stage n**	n0	483
		n1	323
		n2	100

		n3	62
		nx	12
17	**Stage t**	t1	259
		t2	572
		t3	118
		t4	28
		tx	3
18	**Subtype**	HER2+	32
		Luminal_A	392
		Luminal_B	113
		NA	47
		Others	287
		TN	109
19	**Tumor Necrosis Percent**	0	556
		1-10	215
		11-20	103
		21-30	104
		None	2

Table 3: TNBC phenotypic variable categories

3.3.2 SUBGROUPS AND DRUGS ANALYSIS

Patients with triple-negative breast cancer (TNBC), who have negative immunohistochemical staining for estrogen receptors (ER), progesterone receptors (PR), and human epidermal growth factor receptor 2 (HER2) [144], have the lowest survival rate and highest mortality within BC patients [145, 146]. Patients with TNBC have a heterogeneous response to therapy, in which approximately 80% of patients do not completely respond to chemotherapy [147]. The importance of finding better treatments for TNBC has been demonstrated in many studies. Multi-omics data and a network analysis

was used to repurpose drugs for TNBC [27]. The network analysis was used to find modules active in TNBC, but not in normal BC or other BC subtypes. Transcriptional and interactome data were used to create a PPI network to identify highly connected modules and hub genes as candidate drug targets for repurposed therapies [27]. Multi-target drugs were also considered for repositioning to TNBC after creating a drug-target network with the integration of pathways information and the identification of protein-protein interactions [148].

This study offers new guidance in considering heterogeneity within subgroups and the importance of finding subpopulations within TNBC. Moreover, in addition to using the direct relationship between drugs and genes in TNBC, the proposed method leverages other factors, like the perturbation that the drug can cause in the cell. Different criteria were used to filter and select subgroups. The support of the subgroup patterns was greater than 70%, the growth was greater than 1.5, and the confidence was 1. The resulting subgroups were ranked based on their *SPCscore*. The subgroups with an *SPCscore* > 0 were retained for further filtering and analysis because an *SPCscore* = 0 means there is no significant contrast between a given subgroup and the outer population. After filtering the subgroups based on their contrast score and the biomedical importance to our research question, the resulting subgroups were five TNBC subgroups with the patients' neoplasm subdivision, race, and age as clinical variables.

In this work, after analyzing the top drugs (Table 4) for these subgroups, ferroptosis was found as a common cell death mechanism. Ferroptosis is regulated cell death (RCD). It is driven by lipid peroxidation and has an iron dependency [149]. In the subgroups below, more than one drug within the top five drugs play a role in inducing ferroptosis. This result

suggests that focusing on ferroptosis in treating TNBC may be a novel therapeutic route, and may be advantageous for these patients. We found the ferroptosis-inducing drugs were suggested for the subgroups with race and neoplasm subdivision as phenotypic variables. Still, there were no ferroptosis-inducing drugs suggested for the younger subgroups with an age below 50. This finding coincides with the previous finding of ferroptosis being negatively correlated with age [150]. The suggested subgroups are:

a. Subgroup1 is for patients who are Black or African American with no history of a neoadjuvant treatment. This subgroup has 30 patients, and it has 286 unique DEGs to this group. The drugs that can induce ferroptosis that were suggested for this subgroup are Afatinib, Gefitinib, Bosutinib, Lapatinib, and Fulvestrant.
    i. Afatinib was found to be a ferroptosis inducer in TNBC by targeting EGFR [151]. Afatinib's mechanism of action includes binding to wild-type epidermal growth factor receptors (EGFR) and irreversibly inhibiting the kinase activity of all ErbB family members, which are well-recognized as an oncogenic driver in epithelial cancers. Thus, Afatinib can inhibit the proliferation of cancer cell lines [152]. Analyzing the patient profiles in this subgroup showed that EGFR was abnormally expressed in this subgroup. ErbB4 had an abnormal expression as well.
    ii. Gefitinib is an epidermal growth factor receptor (EGFR) and a tyrosine kinase inhibitor. It has antiproliferative and anti-tumoral activity in BC [153, 154]. Gefitinib targets EGFR and is used to induce ferroptosis [151]. The proliferation of TNBC cells can be inhibited by targeting EGFR kinase activity using gefitinib [155]. Moreover, it can be used as a combined therapy with multiple EGFR-TKIs, such as lapatinib and erlotinib, to overcome the resistance to EGFR-targeted therapy in

TNBC. The inhibition of EGFR, in combination with the dual inhibition of cdc7/CDK9, results in reduced cell proliferation, accompanied by the induction of apoptosis, the G2-M cell cycle arrest, the inhibition of DNA replication, and the abrogation of CDK9-mediated transcriptional elongation in TNBC cells [155]. Using gefitinib for this subgroup will inhibit EGFR and cdc7 that was upregulated in this subgroup. Inhibiting them is necessary to reduce cell proliferation. CDK1 and CDK5, which are cyclin-dependent kinases (CDKs), were found to be upregulated in this subgroup. These protein kinases control cancer progression and inhibiting them prevents cancer proliferation [156].

iii. Bosutinib promotes autophagy and research has shown that there is crosstalk between autophagy and ferroptosis [157]. BCR-Abl is the target for bosutinib, dasatinib, imatinib, nilotinib, ponatinib, and regorafenib. Bosutinib is a dual Src/Abl inhibitor [158]. In this subgroup, Src was upregulated; therefore, this drug will inhibit Src and reduce cell proliferation. Src belongs to a different set of mechanisms associated with cancer progression**,** such as inducing a metastatic phenotype, enhancing tumor growth, and enhancing angiogenesis [159].

iv. Lapatinib is a tyrosine kinase inhibitor. This drug was used as a combined therapy with siramesine, which is a lysosome disrupting agent, to induce ferroptosis and reactive oxygen species (ROS) in TNBC [160]. Lapatinib is an inhibitor of ErbB1 and ErbB2, and induces ferroptosis in BC cell lines by altering iron regulation. It is used as a combined therapy to induce ferroptosis in TNBC by inhibiting the iron transport system, leading to an increase in ROS and cell death [161]. Ferroportin-1 (FPN) is an iron transport protein responsible for the removal of iron from cells. Its

expression decreased after treatment with siramesine in combination with lapatinib. The overexpression of FPN decreases ROS and cell death, whereas the knockdown of FPN increases cell death after siramesine and lapatinib treatment. This indicates a novel induction of ferroptosis through altered iron regulation by treating breast cancer cells with a lysosome disruptor and a tyrosine kinase inhibitor [69]. Lapatinib is a tyrosine kinase inhibitor of the epidermal growth factor receptor (EGFR) and ErbB2 (HER2) tyrosine kinases. Studies in vitro showed that lapatinib inhibited the proliferation of cancer cells where the ErbB2 and EGFR are overexpressed [162, 163]. Siramesine and lapatinib gave the best combination index, and this combination induced ferroptosis through iron-mediated ROS and the downregulation of heme oxygenase 1 (HO-1) levels [164]. In TNBC, siramesine and lapatinib increases Transferrin Receptor (TFRC) expression and decreases ferroportin1 (FPN1) expression, thus elevating the level of intracellular iron [165].

b. Subgroup2: The patients in this subgroup are not Hispanic or Latino and White; they have less than three positive lymph nodes, and their histological type is an intraductal carcinoma. This subgroup has 37 patients and 87 unique DEGs that are abnormal in this subgroup, but not in the other subgroups. The drugs that can induce ferroptosis that are suggested for this subgroup are fluvastatin, lovastatin, and gefitinib.
    i. Fluvastatin targets HMG-CoA and can induce ferroptosis by decreasing the expression of glutathione peroxidase 4 (GPX4). This effect is time- and concentration-dependent [166, 167]. Treatment with this statin induces ferroptosis by inhibiting GPX4 and the key products in the mevalonate pathway, like 3-hydroxy-3-methyl-glutaryl-coenzyme A reductase, and the depletion of CoQ10, that reduces

the levels of this key membrane antioxidant [166, 168, 169]. Fluvastatin can be used as a combined therapy with RSL3, which is a direct inhibitor of GPX4 [166].

ii. Lovastatin: Lovastatin is another statin drug that can inhibit GPX4 and induce ferroptosis [170]. Similar to fluvastatin, the lovastatin target point is lipid synthesis, and it represents a HMGCR/HMG-CoA reductase inhibitor [171, 172].

iii. Gefitinib: As mentioned previously, gefitinib can be used to induce ferroptosis in TNBC [60].

After analyzing the genes that are affected by these drugs in this subgroup, we found that fluvastatin, lovastatin, and gefitinib downregulate CCT5, which was upregulated in this subgroup. CCT5 interacts with HMGCR, NFE2L2, and TP53. In the TP53 pathway, the upregulation of GLS2, which is a transcription target of TP53, leads to P53-dependent ferroptosis, and the inhibition of SC7A11 by TP53 can also trigger ferroptosis. TP53 (tumor protein 53) represses SLC7A11 to promote ferroptosis as a tumor suppression mechanism [173] (SLC7A11 was upregulated in this subgroup). Fluvastatin and lovastatin bind to HMGCR and upregulate CDKN1A, which is a ferroptosis regulation gene [150].

c. Subgroup3: The patients in this subgroup are not Hispanic or Latino. They have right sided TNBC as their neoplasm subdivision, they have less than three lymph nodes that are positive, and their histological type is an intraductal carcinoma. This subgroup contains 30 patients and has 118 unique DEGs that are abnormal in this subgroup, but not in other subgroups. The drugs that can induce ferroptosis that are suggested for this subgroup are sunitinib and pazopanib.

i. Sunitinib: Sunitinib can induce ferroptosis by targeting the von Hippel-Lindau (VHL), which increases sensitivity to ferroptosis. Sunitinib interacted with 48 DEGs in this subgroup, including CDO1 [174]. The inhibition of CDO1 increases ROS and can induce ferroptosis [175].

ii. Pazopanib: Pazopanib belongs to the same category as Sunitinib. Moreover, treating breast cancer cells with Pazopanib was found to induce autophagic cell death [176]. It is a necroptosis inhibitor [177]. The mTOR pathway is a regulator of iron metabolism. The VHL/HIF-α axis is the main regulator target of iron metabolism. HIF-3α was downregulated in this subgroup. Targeting the VHL gene pathway using drugs like sunitinib, sorafenib, pazopanib, and axitinib causes the VHL to be inactive. The inactivation of the VHL increases sensitivity to ferroptosis. IRP1 can bind to the iron reaction element of HIF-2α mRNA and inhibit its translation. Tempol, an IRP1-activated drug, inhibits HIF-2α and HIF-1α protein levels. PT2399 and PT2385 are inhibitors of HIF-2α [178]. Pazopanib is a tyrosine kinase inhibitor (TKI) that belongs to a class of drugs that targets PDGFR α/β and VEGFR activity [158]. PDGFR was found to be downregulated in this subgroup.

After analyzing the genes in this subgroup, we found that sunitinib and pazopanib bind to FGFR2, which, in turn, interacts with HSPB1 and STAT3. FGFR2 was upregulated in this subgroup. Both genes play a role in ferroptosis induction, as explained previously. Moreover, sunitinib binds to PHKG2, which is important for the induction of ferroptosis [179-181].

d. Subgroup4: The patients in this subgroup are not Hispanic or Latino; they have left side TNBC as their Neoplasm Subdivision. They have less than three lymph nodes that are

positive, and their histological type is an intraductal carcinoma. This subgroup has 32 patients and has 103 unique DEGs that are abnormal in this subgroup, but not in the other subgroups. The drugs that have the potential to induce ferroptosis in this subgroup are dexamethasone and vorinostat.

i. Dexamethasone: High-dose dexamethasone disrupts the metabolism of glutamate and cysteine, produces more ROS, and downregulates GPX4 and system XC−, which are two key mediators of ferroptosis [182]. In this subgroup, ROS1 was upregulated, GPX2-3 were downregulated, GPX5 was upregulated, and GPX4 was in the normal range.
ii. Vorinostat: Vorinostat is a histone deacetylase (HDAC) inhibitor. In colon cancer cells, vorinostat can significantly inhibit cell growth and can induce cell cycle arrest and apoptosis [183]. It may induce ferroptosis by glutamine deprivation resulting in the accumulation of ROS [184]. In NSCLC, vorinostat combined with erastin, can increase the lipid peroxide levels and can inhibit HDAC to induce ferroptosis [185]. Moreover, when vorinostat is used as a combined therapy with gefitinib or erlotinib, which are EGFR-TKIs, it can promote oxidative stress-dependent apoptosis by the suppression of the c-MYC-regulated NRF2 functions and increase the levels of KEAP1 in NSCLC cells [186].

The analysis of this subgroup's genes and drugs showed that dexamethasone bound to PTGS2, downregulated NFE2L2, and upregulated HSPB1, STAT3, and CDKN1A. These kinds of regulations promote ferroptosis. Dexamethasone downregulates GDF15 and NFKB2, and it upregulates MEGF9. In this subgroup, GDF15 was upregulated, NFKB2 was upregulated, and MEGF9 was downregulated. GDF15 interacts with

ferroptosis-related genes like ATG7, EGFR, FTH1, HSPB1, PHKG2, PTGS2, AKR1C2, CDKN1A, DPP4, NFE2L2, and TP53. In this subgroup, AKR1C2 and DPP4 were downregulated. NFKB2, which was upregulated in this subgroup, interacted with EGFR, CDKN1A, and HSPA5. MEGF9, which was downregulated in this subgroup, interacted with GCLM, HMOX1, PTGS2, AKR1C2, which were downregulated in this subgroup; FDFT1 and NFE2L, which were upregulated in this subgroup; and GCLM. The downregulation of FTH1 can induce ferroptosis. IRP1 and IRP2, which were downregulated in this subgroup, are iron regulatory proteins that can regulate iron metabolism genes such as TFRC and FTH1 to maintain the stability of LIP [173]. In this subgroup, TFRC was upregulated while LIPC and LIPE were downregulated. AKR1C2, which was downregulated in this subgroup, is one of the ferroptosis regulators. AKR1C1-3 (aldo-keto reductase family 1 member C1), which was downregulated in this subgroup, is involved in eliminating end products of lipid peroxidation [150]. Vorinostat upregulates HMGCR, HMOX1, PHKG2, CDKN1A, and FDFT1.

e. Subgroup5: The patients in this subgroup are less than 50 years old; their pathological stage is m0 with no history of neoadjuvant treatment. This subgroup has 37 patients and has 216 unique DEGs that are abnormal in this subgroup, but not in other subgroups. Contrary to what was observed in the other four subgroups, most of the top drugs that are suggested to this subgroup have an antioxidative effect on cancer cells. These drugs are rifampicin, galantamine, cerulenin, and lipoic acid.
    i. Rifampicin: Rifampicin was repurposed as an antiferroptosis agent, and it functions as a lipid peroxyl radical scavenger [187, 188].

ii. Galantamine: Galantamine has an antioxidant effect and acethylcholinesterase and γ- secretase inhibitory activity [189].

iii. Cerulenin: Cerulenin is an antifungal antibiotic that inhibits fatty acid and steroid biosynthesis. Fatty acid synthase (FAS) was observed to have a high expression in breast cancer cells in comparison with normal cells, and there is increasing evidence that FAS plays a role in breast cancer development [190]. Moreover, fatty acid synthase (FAS) and ErbB2 have been shown to promote breast cancer cell migration [191]. The effect of cerulenin on breast cancer was tested in vivo analysis. Due to cerulenin's ability to inhibit fatty acid synthase (FAS) that, in turn, decreases the expression of ErbB1, 2, and 4, it may initiate an epithelial-to-mesenchymal transition

ID	Subgroup1	Subgroup2	Subgroup3	Subgroup4	Subgroup5
1	Digitoxin	Homoharringtonine	Sunitinib	Idarubicin	Rifampicin
2	Ouabain	Chenodeoxycholic acid	Pazopanib	Topotecan	Varenicline
3	Digoxin	Fluvastatin	Mycophenolic acid	Mitomycin	Tubocurarine
4	Daunorubicin	Lovastatin	Lurasidone	Vinblastine	Cerulenin
5	Afatinib	Gefitinib	Mianserin	Gemcitabine	Galantamine
6	Bosutinib	Bosutinib	Mirtazapine	Tolazamide	Sorafenib
7	Lapatinib	Erlotinib	Risperidone	Levetiracetam	Topotecan
8	Gefitinib	Cyclosporine	Asenapine	Epirubicin	Vemurafenib
9	Tretinoin	Sunitinib	Cabergoline	Irinotecan	Methylphenobarbital
10	Mebendazole	Mycophenolic acid	Iloperidone	Nilotinib	Primidone

Table 4: Top ten drugs for each of the five TNBC subgroup interest

(EMT) as well as the migration and invasive ability of cancer cells [192].

iv. Lipoic acid (LA): Lipoic acid is an antioxidant. It has an anti-proliferative effect in breast cancer cells by reducing breast cancer cell viability, cell cycle progression, and the EMT. It downregulates furin, which, in turn, inhibits the maturation of IGF-1R [193, 194]. Combining lipoic acid with radiation therapy was found to overcome the resistance to radiation therapy and promote apoptosis [195].

3.3.3 PATHWAY ENRICHMENT ANALYSIS

In order to determine common mechanisms underlying the effects of the most effective drugs for each of the five subgroups, a pathway enrichment analysis was performed (Figure 10). After removing the duplication of drugs between subgroups, 46 unique drugs in total were identified for the five subgroups. An enrichment analysis for pathways targeted by these drugs was performed. The statistical significance of these pathways was calculated using Enrichr [87, 196, 197] with $p < 0.05$. The metabolism pathway was targeted by 100% of the drugs. Metabolic reprogramming is considered to be a hallmark of cancer that can be used for diagnosis, prognosis, and treatment [198]. Cancer cells show different metabolism patterns as compared with normal cells, and targeting that difference is a promising strategy for developing anticancer therapy [199]. Studies have shown that targeting metabolic enzymes can reverse drug resistance in cancer [200, 201]. Targeting metabolism pathways could be an important component in developing a comprehensive treatment for breast cancer, considering the complexity of these pathways due to their crosstalk with other signaling pathways [202]. The upregulation of some metabolic pathways was found to stimulate metastasis in breast cancer [203]. The importance of targeting metabolism pathways also has been shown clearly in TNBC, where there is a metabolic heterogeneity between patients of this cancer subtype. The metabolism pathways were used to stratify TNBC patients into subgroups based on the heterogeneity of the metabolism [204, 205]. Our findings coincide with what has been suggested in the literature and points to the importance of developing drugs that target metabolism pathways to tailor therapies for TNBC patients. The metabolic heterogeneity of TNBC was targeted by the top drugs, those directed at specific gene patterns contained within these subgroups.

These patterns belong to the metabolism, but they were unique to a given subgroup and had high contrast with the outer population.

The second pathway, targeted by 93% of the top drugs, was the biological oxidation pathway. Increased oxidative stress plays a role in carcinogenesis in breast cancer patients [206]. In a study of the immune microenvironment of breast cancer, oxidative stress in combination with other biological phenotypes was associated with high immune infiltration [207]. Thus, targeting oxidation pathways may be promising for developing immune therapies for breast cancer. Studies have shown that the spread of BC metastatic cells and their survival in circulation depends on different factors, including antioxidant protection [208]. The oxidative stress response is used by breast cancer cells to adapt to their nutrient-poor environment [209]. Oxidative activity is one factor involved in the induction of breast carcinogenesis [210].

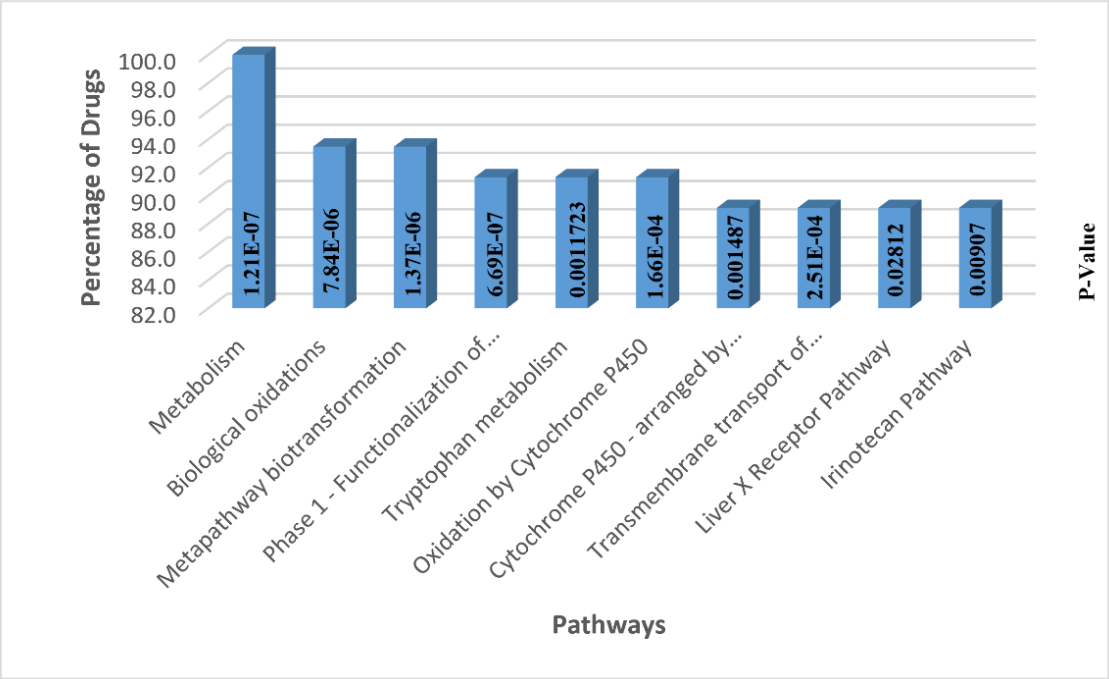

Figure 10: Pathways enrichment analysis for the genes targeted by top ten drugs for all the five subgroups of interest. The x-axis shows pathway names, and the y-axis shows the percentage of drugs that target each of these pathways.

The metabolism and oxidation-related pathways can be observed more frequently than others when considering the pathways targeted by the top drugs. In analyzing the drugs, the focus was on those shared a common mechanism within the top five drugs. Still, the pathway enrichment analysis showed that the top ten drugs of all subgroups targeting metabolism and oxidation-related processes inhibit cancer growth and development. Moreover, an enrichment analysis for GO domains, cellular components, biological processes, and molecular functions was done as a part of this study. The top molecular functions that were highly enriched for the top ten drugs were oxidoreductase activity, protein homodimerization activity, and lipid binding. (Figure 11). The top biological process enriched for the top drugs were the response to the drug, the oxidation-reduction process, and the response to the organonitrogen compound (Figure 12). The top cellular components were the endoplasmic reticulum membrane, nuclear outer membrane-endoplasmic reticulum membrane network, and plasma membrane region (Figure 13).

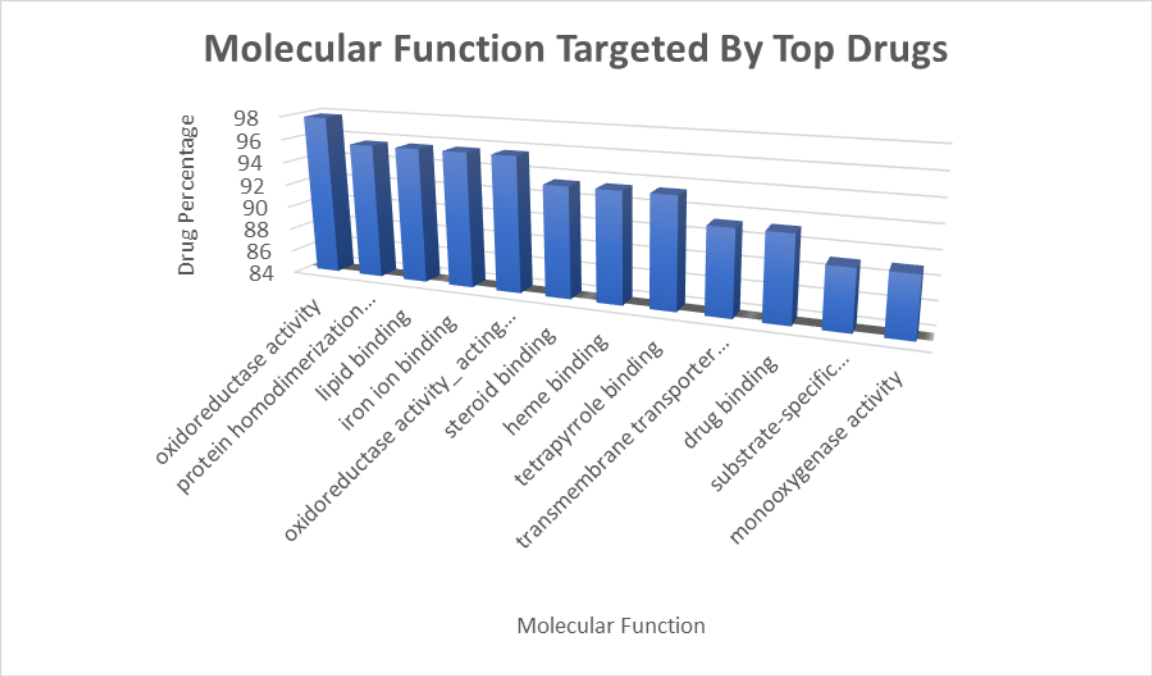

Figure 11: Top molecular functions targeted by top ten drugs of the five subgroups of interest.

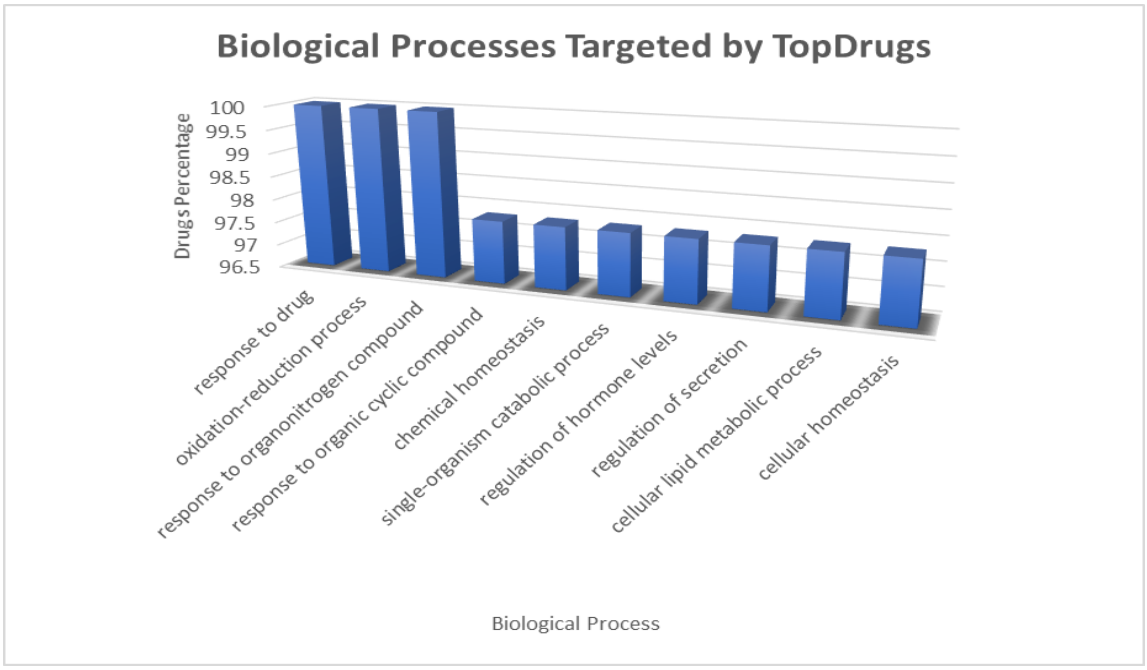

Figure 12: Top biological processes targeted by top ten drugs of the five subgroups of interest.

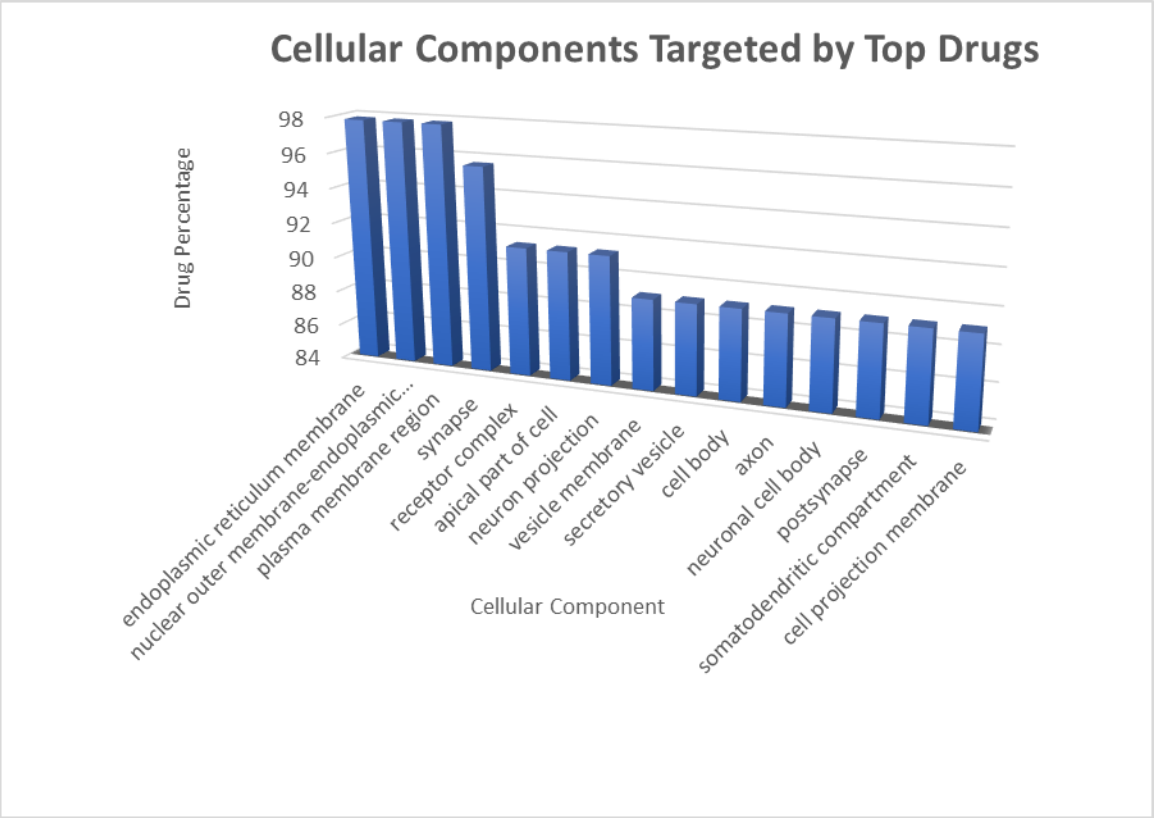

Figure 13: Top cellular components targeted by top ten drugs of the five subgroups of interest.

3.3.4 PHARMACOLOGICAL CLASSES ANALYSIS

A drug class enrichment analysis was done for the top ten drugs of the five subgroups of interest to find the pharmacological classes to which most of the drugs belong. Using the FDA Established Pharmacologic Class (EPC), classes for 32 drugs were found. The top pharmacological class to which most of these drugs belong was the kinase inhibitor. Kinase inhibitors have shown great potential as a TNBC regimen due to their activity as antitumor and antiangiogenic therapies by targeting genes and pathways that play a role in cancer development and proliferation, like targeting EGFR, RON, MET, mTOR, BRAF, MEK, Src, and Bcr/Abl [211-216]. Kinase inhibitor drugs have been used in different breast cancer research and clinical trials, especially for TNBC, and they are

considered to be one of the most successful targeted therapies for cancer [217, 218]. They have been used as a monotherapy and combined therapy for TNBC, and they are considered a promising strategy for treating advanced TNBC [219]. As a monotherapy, a tyrosine kinase inhibitor effectively inhibits the proliferation and induced autophagy and apoptosis in TNBC cells [220]. A kinase inhibitor was found to have the ability to preferentially suppress the growth and proliferation of TNBC cells in comparison with non-TNBC cells [221]. It was found to reduce the proliferation of mesenchymal stem-like (MSL) cells and induce apoptosis in TNBC cells as a combined therapy [222, 223]. Moreover, a combination of kinase inhibitors was suggested for TNBC, where monotherapy has limited success [224]. Kinase inhibitors were used in combination with other inhibitors to overcome the resistance for monotherapy, like using dasatinib in combination with Akt or mTOR inhibitors to overcome dasatinib resistance and reduce the proliferation in TNBC cells [225]. Many kinase inhibitors were effective as a TNBC therapy, and many others are a subject of ongoing clinical trials [226, 227].

3.3.5 DISCUSSION

Breast cancer is a highly heterogeneous disease. Its heterogeneity requires regimens to be tailored for homogeneous subgroups of patients with common phenotypic and genotypic features that are unique to each subgroup. The data-driven approach developed in this study was implemented to stratify breast cancer into subgroups using their phenotypic and genotypic data. The method began with the phenotypic features and partitions the disease population based on the categorical values of the phenotypic variables. This partitioning created different subgroups. For each subgroup, the algorithm used the genotypic features

to evaluate druggability potential, which refers to the necessity of treating a given subgroup differently and tailoring drugs to that subgroup.

Each subgroup's genotypic features are represented as a network and processed using a graph database, neo4j. Each network consists of heterogeneous biomedical entities. The core of each network is the differentially expressed gene patterns in that subgroup. For each gene, the network contains other genes that directly interact with that gene, the pathways and GO domains to which that gene belongs, and the drugs that affect the expression of the other genes in that network. The opposite expression regulation between the drugs and the affected genes was considered by retrieving the drugs from the knowledgebase into the subgroup's network using cMAP data. Each network has several drugs, and to find the best drugs that can be suggested for a given subgroup, the comprehensive drug scoring algorithm was implemented within each network using different factors. These factors include the percentage of differentially expressed genes, the number of patterns targeted by each drug, and the perturbation impact of a drug on human cells.

Breast cancer has three major subtypes, luminal A, luminal B, and triple-negative breast cancer (TNBC). TNBC patients have the lowest survival rate among all breast cancer patients. This was the motivation to focus the study on TNBC patients. Five subgroups were analyzed, and these subgroups composed of combinations of the three population variables, which are age, race, and neoplasm subdivision, as statistically significant phenotypic variables. After analyzing the differentially expressed genes for these subgroups, it was found that these subgroups had, on average, 162 unique differentially expressed genes in addition to phenotypic-level differences. This uniqueness does not only

mean the gene is differentially expressed in a given subgroup, but it also could mean the gene has a differential expression pattern unique to a given subgroup. It could be only upregulated in that subgroup while it is downregulated in the other subgroups, and vice versa.

Analyzing the top five drugs for the subgroups of interest showed a common mechanism shared by four subgroups, with race and neoplasm subdivision as the significant phenotypic factors. Still, the opposite mechanism was suggested for the subgroup with age as a significant phenotypic factor. Inducing ferroptosis was the common mechanism in the subgroups that were Black or African American, were White, had right side breast cancer, or the subgroups with left side breast cancer. In contrast, the antioxidant was the common targeted mechanism in the subgroup with patients aged less than 50 years. This finding coincides with what has been found previously [150], where age is negatively correlated with ferroptosis. The pathway enrichment analysis also showed that the most targeted pathways by the top ten drugs were pathways that have a relation to metabolism and oxidation. The kinase inhibitor is considered one of the most targeted therapies to treat

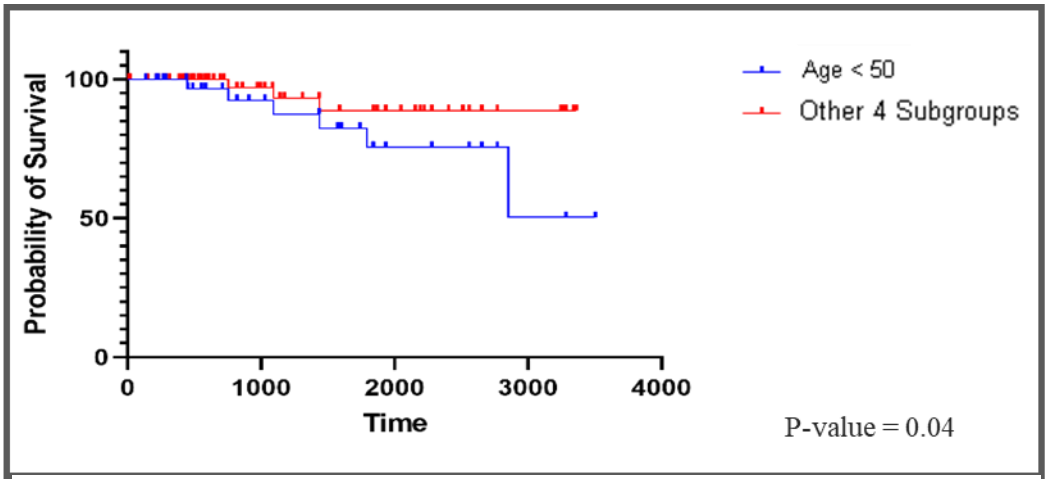

Figure 14: Survival curves for subgroup5 vs. other subgroups. The curve shows the patients younger than 50 who have lower survival rates than the other 4 subgroups of interest.

breast cancer, especially TNBC. Since it is a computational-based method, a literature review was used to validate the findings.

The clinical relevance of the data concerning repurposing TKIs as ferroptosis inducers is strengthened by recent findings based on the functional analysis of breast cancer cell lines that demonstrate that TNBCs are enriched in ferroptosis gene signatures and are vulnerable to ferroptosis inducers, compared to other breast cancer subtypes, i.e., ER-positive and/or HER2-positive subtypes [228]. Furthermore, age is the most pertinent clinical characteristic that has been shown to have a negative correlation with ferroptosis in certain tumors like BC, i.e., in younger BC patients, ferroptosis is less clinically relevant [229] which coincides with our discovery of a completely different class of repurposed medications that did not induce ferroptosis in the younger (< 50 years) cohort of BC patients. The potential benefit from antioxidants in younger TNBC patients is further supported by biological data on the dependence on reactive oxygen species (ROS) for the survival and malignant progression of TNBC in younger patients. These high levels of ROS production that drive the aggressiveness in TNBC resulting from oncogenic gene mutations, e.g., BRCA1 [230], gene expression changes, e.g., BLT2 [231] and the attainment of stem-cell like properties [232]. This high ROS content induces multiple signaling which, in turn, leads to highly proliferative, migratory, and drug-resistant phenotypes in TNBC and is effectively attenuated by antioxidants, thereby reducing the growth and metastasis of TNBC cells [233]. The shorter survival observed in the younger cohort or subgroup 5 when it was compared against all remaining four subgroups “bundled” together (Figure 14) is a testament to the different ROS-dependent biology in

subgroup 5, which contributes to the cell survival, metastasis, chemoresistance, and cancer relapse. Therefore, the additional use of antioxidants can be an effective strategy for treating these younger TNBC breast cancer patients in order to prevent chemoresistance through depleting ROS.

Still, these findings need more testing in a wet lab setting to ensure they are valid to be suggested for healthcare improvement. After validation in the wet lab, these results can significantly improve TNBC survival and suggests drugs tailored for a specific group of patients based on their phenotypic and genotypic features.

3.4 PAN-CANCER ANALYSIS RESULTS

Cancer is a group of diseases that is the second leading cause of death in the United States. Each cancer type has a different mortality rate, and patients of each type have heterogeneous responses to treatment. However, all these types of cancer are characterized by uncontrolled growth and the spread of abnormal cells. This suggests the existence of common mechanisms among these types in addition to the unique mechanisms for each type. Studying the inter and intra heterogeneity of cancer is crucial to understanding the mechanisms of action, identifying biomarkers, finding drug targets, and developing or repurposing therapies. To study the heterogeneity of cancer, pan-cancer analyses need to be conducted over a wide range of cancer types. The purpose of these analyses is to find homogeneous subgroups of patients across different cancer types. Finding these subgroups enables targeting the cancer mechanism over different cancer types and this can reduce the cost of treating patients because one drug can be used as a regimen for more than one cancer type. Additionally, finding subgroups across cancer types will improve patient

survival by finding a better treatment through targeting common mechanisms instead of only targeting a mechanism unique to a specific cancer type. Also, to reduce the cost associated with treating patients, old drugs can be investigated for new uses by developing and implementing drug repositioning methods over pan-cancer data after stratifying patients into subgroups.

Pan-cancer represents a comprehensive heterogeneity analysis required to solve the intra-heterogeneity problem which is the major barrier to classifying patients into potential benefited groups [234]. This type of analysis has been used for a variety of research questions including studying genes' effect on cancer in general instead of studying their effect on each cancer type [235, 236]. For the subgrouping, pan-cancer analysis has been done using different methods to stratify cancer patients into subgroups. It was used to stratify patients based on the expression status of one gene and its upstream, downstream, and correlated genes to study cancer prognosis in each subgroup [237, 238]. This type of study does not consider the wide range of the genetic variation because it focuses on a narrow set of genes. A wider set of molecular features were considered for stratifying patients into groups where patients in each subgroup share the same molecular features. These molecular features were represented by RNA signatures, Tumor Mutational Burden (TMB), Copy-Number Alteration (CNA), and genes expression using network analysis [234, 239-241]. These methods represent an improvement on previous methods that depend on a limited set of genes, but they only focus on genotypic features without taking into account heterogeneity on the phenotypic level. Other methods addressed the importance phenotypic heterogeneity by using the phenotypic features to stratify patients into subgroups. For the stratification purposes, the surgery and radiotherapy status, socio-

economic status, mortality after surgery, race, age, and metastasis site were used [242-244]. While these methods do use the phenotypic features, they miss the importance of genotypic features to stratify patients into subgroups.

Pan-cancer analysis has been used to identify novel drug targets that can be advantageous for different types of cancer [245]. One gene-based analysis was developed to identify drugs that can be repurposed based on their ability to switch the expression of a gene [104]. Using "one gene, one drug" is not an effective method to treat a heterogeneous disease like cancer. Pan-cancer analysis was used to find novel cancer biomarkers and drug targets that can be useful for different cancer types and used for drug repositioning using more than one gene including multi-omics data [245-248]. Also, network analysis was used to include more factors that can affect cancer prognosis and development like using multi-omics data with pathway and pharmacogenomic data to reposition adrug [249-251]. These methods used a wide range of biomedical data to repurpose drugs, but they analyzed the pan-cancer data as whole without addressing the importance of identifying homogeneous subgroups within this heterogeneous cancer population.

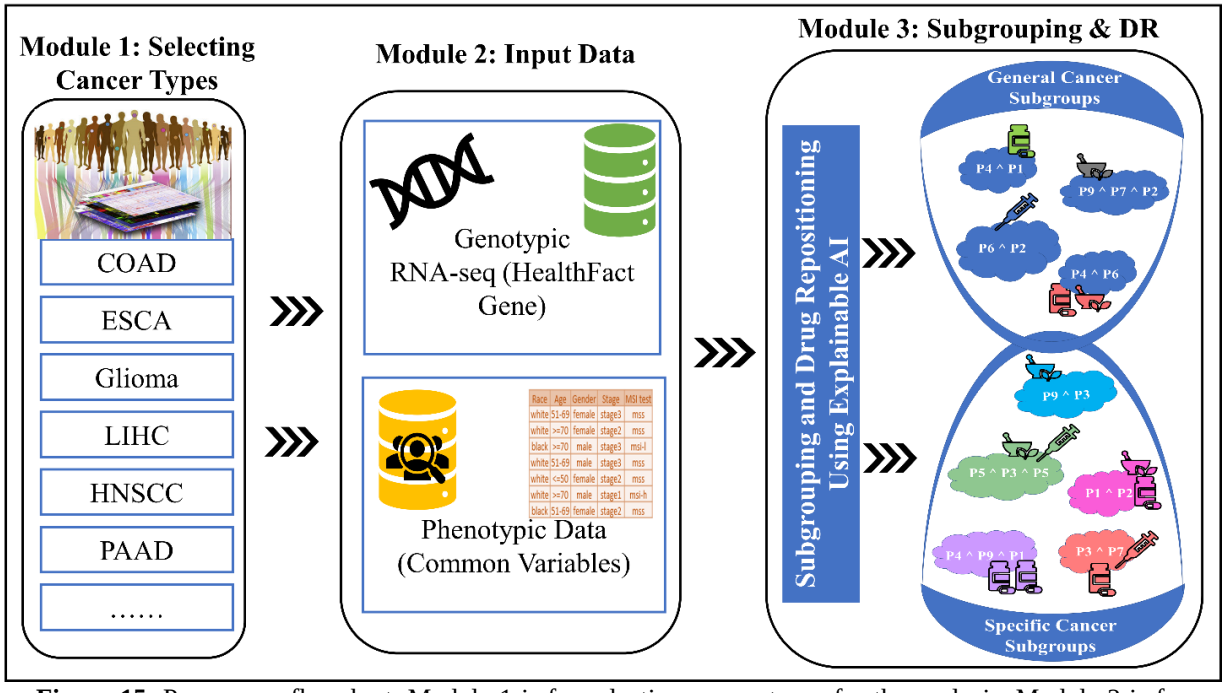

Figure 15: Pan-cancer flowchart. Module 1 is for selecting cancer types for the analysis. Module 2 is for obtaining the phenotypic and genotypic data from TCGA for the patients with the selected cancer types. Module 3 is for the implementation of the subgrouping and drug repositioning method to stratify patients into subgroups, and the subgroups divide into general subgroups and specific cancer subgroups. The general subgroups share subgroup phenotypic variables among more than one cancer type, while the cancer-specific subgroups have unique phenotypic variables to each cancer type.

There is an urgent need to develop and implement computational methods that stratify patients into subgroups and reposition drugs for each subgroup based on phenotypic features and a wide range of genotypic features that capture a broader range of molecular assortment. In this part of the study, a pan-cancer analysis was conducted to stratify patients into homogeneous subgroups using phenotypic and genotypic data. Then, the drug repositioning method was implemented to reposition drugs for each subgroup. The explainable AI subgrouping and drug repositioning method was implemented to stratify patients into subgroups and recommend drugs to be repurposed for each subgroup of patients. The algorithm identified subgroups that are common across cancer types and unique to each cancer type and recommended drugs for each subgroup.

This section consists of three subsections. In section 3.4.1, the data description and processing were presented (Figure 14-Module 1 and 2). Section 3.4.2 is the patient stratification results, where the subgroups finding, and selection was described (Figure 14-Module 3). In section 3.4.3, the drug repositioning results are explained and analyzed (Figure 14-Module 3).

3.4.1 DATA DESCRIPTION AND PROCESSING

In this analysis, the total number of patients was 3983. Fourteen caner types and one subtype in 11 organs were included (Table 5). selection of the cancer types was based on the survival percentage that was documented in The Surveillance, Epidemiology, and End Results (SEER) data and the number of available cases in TCGA. The survival threshold was 75%, where cancer types that have less than 75% survival rate were selected for this analysis. Out of these cancer types, the diseases that have data for more than 100 patients in TCGA were kept. The dataset for these patients consists of genotypic and phenotypic variables. It has 24 phenotypic consist of mutation status for 14 genes that were identified as cancer drug targets and 10 clinical variables (Table 6). The genotypic data consists of 2415 genotypic variables which are the RNA-seq data of the cohort. The continuous variables in the phenotypic dataset were categorized based on the medical guidelines and the physician recommendations. The genotypic variables were categorized based on the z-score of each gene in that dataset.

ID	Location	Pathology Type	#Cases	#Disease
1	Brain	Glioma (GBM & LGG)	159	2
2	Breast	Triple Negative Breast Cancer (TNBC)	109	1
3	Colon	Colon Cancer (COAD)	214	1
4	Esophagus	Esophagus (ESCA)	184	1
5	Kidney	Pan-kidney cohort (KICH & KIRC & KIRP)	889	3
6	Liver	Liver hepatocellular carcinoma (LIHC)	371	1
7	Lung	Lung adenocarcinoma & Lung squamous (LUAD & LUSC)	1007	2
8	Ovary	Ovarian serous cystadenocarcinoma (OV)	300	1
9	Pancreas	Pancreatic adenocarcinoma (PAAD)	178	1
10	Rectum	Rectum Cancer (READ)	182	1
11	Stomach	Stomach and Esophageal carcinoma (STAD & ESCA)	378	2

Table 5: Cancer types included in the analysis

	Phenotypic Variable	Categories
1	Disease code	GBM, LGG, LUSC, KIRC, LUAD, OV, HNSC, …
2	Age	<50, 50-65, 65+
3	Ethnicity	Not Hispanic or Latino, Hispanic or Latino
4	Gender	Male, Female
5	Race	Black Or African American, White, Asian
6	Pathologic Stage	Stage I, Stage ii, Stage iii, Stage iv
7	Stage Categories m	m0, m1, mx
8	Stage Categories n	n0, n1, n2, n3, nx
9	Stage Categories t	t1, t2, t3, t4
10	Tumor Location	Brain, Breast, Colon, Esophagus, Kidney, Liver, Lung, Ovary, Pancreas, Rectum, Stomach
11	Mutation data	SRC, FOS, TP53, P53, PCL2, KRAS, BRAF, EGFR, BRCA1, BRCA2

Table 6: The Phenotypic variables for pan-cancer analysis

3.4.2 PATIENT STRATIFICATION RESULTS

The implementation of the subpopulation discovery process (Section 2.3) resulted in a set of subgroups with a contrast score for each (SPC_{Score}). The subgroups with significant contrast from the outer population were retained for further analysis and finding repositioning drug candidates. The subgroups were initially analyzed based on the phenotypic variables' distribution across cancer types. The subgroups with gender as a significant phenotypic variable were widely shown in most cancer types (Figure 15). So, these subgroups were considered to be the focus subgroups or subgroups of interest (SOI) for this analysis. Three hundred and one subgroups have gender as a phenotypic variable divided into 108 female subgroups and 193 male subgroups.

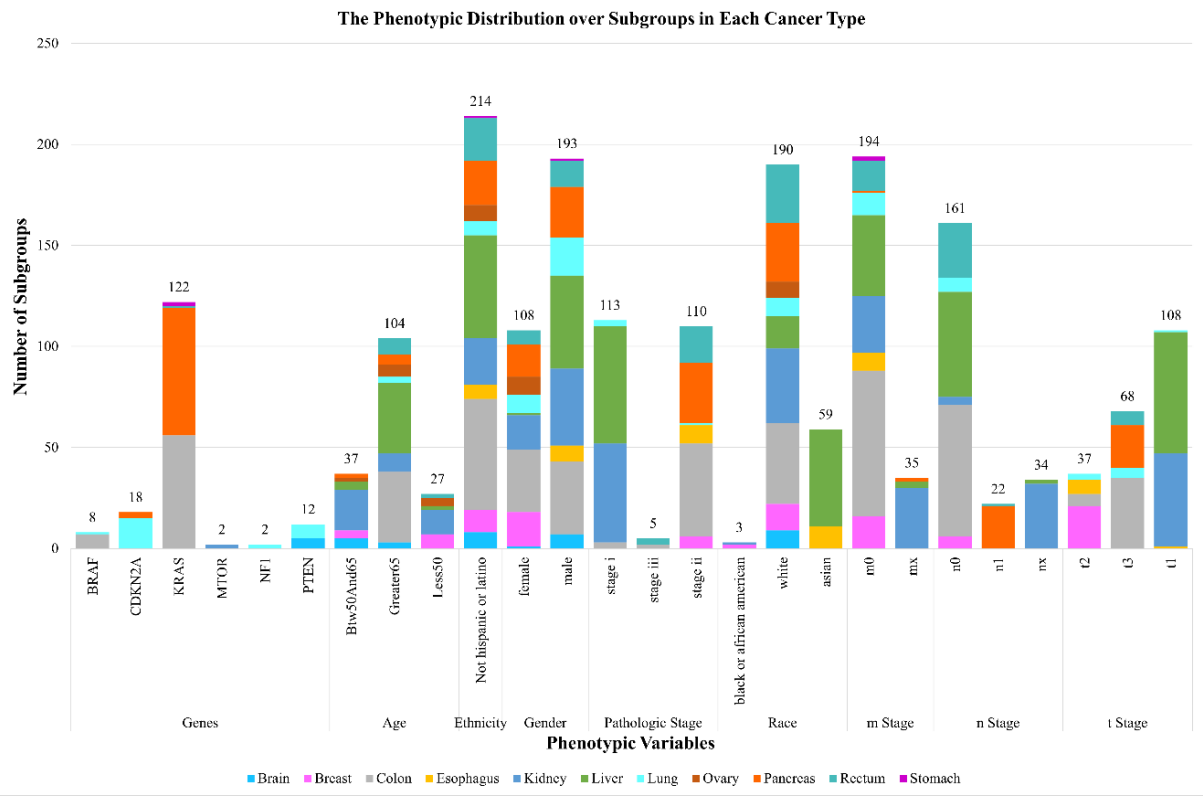

Figure 16: Subgroup phenotypic variables distribution over cancer types

These subgroups were clustered based on phenotypic variable similarity regardless of cancer types to find similar subgroups across cancer types. For example, let's SG1 be a female subgroup with (P1^P2^D1^F) as its phenotypic variables, and SG2 is a male subgroup with (P1^P2^D2^M) as its phenotypic variables. These two subgroups were clustered together because they have the same phenotypic (clinical) variables (P1 and P2) and the only difference besides the gender (F=female in SG1, and M=male in SG1) is the cancer type (D1 is the disease's type of SG1 while D2 is the disease's type of SG2). This clustering is to find the genotypic and phenotypic characteristics that are true across varying cancer types. Finding these characteristics will help uncover candidate drugs that can work for specific subgroups of patients across cancer types. After implementing this clustering, there were 33 female subgroups and 39 subgroups with a male as a significant phenotypic feature. These subgroups vary in the number of cancer types. The focus was on the subgroups that appeared in more than one cancer type. There were 25 female (Table 7) and 25 male (Table 8) subgroups with the same phenotypic variables but appeared in different types of cancer. Five subgroups were common between the gender types. These subgroups appeared in different types of cancers for both male and female, 13 female subgroups and 13 male subgroups (Top 5 subgroups in Table 7 & 8). These subgroups were selected to be studied in more detail in the context of genotypic features used to recommend drugs and to analyze the biological mechanisms of these subgroups. These subgroups were the input to the next step, drug repositioning, to find the candidate drugs for each subgroup across cancer types.

Female Subgroups		
ID	**Subgroup**	**Cancer Cite**
1	Age=greater65 & Gender=female	colon, lung, ovary
2	Gender=female & Race=white	brain, lung, ovary
3	Gender=female & m.Stage=m0	colon, kidney, lung
4	Gender=female & Mutation=KRAS	colon, pancreas
5	t.Stage=t1 & Gender=female	kidney, lung
6	n.Stage=n0 & m.Stage=m0 & Gender=female & Race=white	colon
7	Age=greater65 & Gender=female & n.Stage=n0 & m.Stage=m0	colon
8	Gender=female & Mutation=KRAS & Stage=stageii	pancreas
9	Gender=female & Mutation=KRAS & Race=white	pancreas
10	n.Stage=n0 & m.Stage=m0 & Gender=female & Race=white & Ethnicity=not_hispanic_or_latino	colon
11	Mutation=KRAS & Gender=female & Race=white & t.Stage=t3	pancreas
12	Mutation=KRAS & Gender=female & t.Stage=t3	pancreas
13	Age=btw50and65 & Gender=female	ovary
14	Age=greater65 & Gender=female & n.Stage=n0	colon
15	Age=greater65 & m.Stage=m0 & Gender=female	colon
16	Gender=female & Mutation=KRAS & m.Stage=m0	colon
17	Gender=female & Mutation=KRAS & n.Stage=n0	colon
18	Gender=female & Mutation=KRAS & n.Stage=n1	pancreas
19	Gender=female & Stage=stageii	lung
20	Stage=stageii & Gender=female & Ethnicity=not_hispanic_or_latino & t.Stage=t3	colon
21	Stage=stageii & m.Stage=m0 & Gender=female & t.Stage=t3	colon
22	m.Stage=m0 & Gender=female & t.Stage=t3	colon
23	n.Stage=n1 & Stage=stageii & Mutation=KRAS & Gender=female	pancreas
24	t.Stage=t1 & Gender=female & Race=white & Stage=stagei	kidney
25	t.Stage=t3 & Gender=female & Mutation=KRAS & Stage=stageii	pancreas

Table 7: Female subgroups of interest for pan-cancer analysis

Male Subgroups		
ID	**Subgroup**	**Cancer Cite**
1	Age=greater65 & Gender=male	brain, colon, kidney
2	Gender=male & Race=white	kidney, lung, rectum
3	Gender=male & m.Stage=m0	colon, liver, lung
4	Gender=male & Mutation=KRAS	colon, pancreas
5	Gender=male & t.Stage=t1	kidney, liver
6	n.Stage=n0 & m.Stage=m0 & Gender=male & Race=white	colon, liver, rectum
7	Age=greater65 & Gender=male & n.Stage=n0 & m.Stage=m0	colon, liver
8	Gender=male & Mutation=KRAS & Stage=stageii	colon, pancreas
9	Mutation=KRAS & Gender=male & Race=white	colon, pancreas
10	n.Stage=n0 & m.Stage=m0 & Gender=male & Race=white & Ethnicity=not_hispanic_or_latino	colon, rectum
11	t.Stage=t3 & Gender=male & Mutation=KRAS & Race=white	pancreas
12	Mutation=KRAS & Gender=male & t.Stage=t3	pancreas
13	Age=btw50and65 & Gender=male	brain
14	Age=greater65 & n.Stage=n0 & Gender=male	colon
15	Age=greater65 & m.Stage=m0 & Gender=male	liver
16	Gender=male & Mutation=KRAS & m.Stage=m0	colon
17	n.Stage=n0 & Mutation=KRAS & Gender=male	colon
18	n.Stage=n1 & Mutation=KRAS & Gender=male	pancreas
19	Gender=male & Stage=stageii	colon
20	t.Stage=t3 & Gender=male & Ethnicity=not_hispanic_or_latino & Stage=stageii	colon
21	Stage=stageii & m.Stage=m0 & Gender=male & t.Stage=t3	colon
22	t.Stage=t3 & Gender=male & m.Stage=m0	lung
23	n.Stage=n1 & Stage=stageii & Mutation=KRAS & Gender=male	pancreas
24	Stage=stagei & Gender=male & Race=white & t.Stage=t1	liver
25	Stage=stageii & Mutation=KRAS & Gender=male & t.Stage=t3	pancreas

Table 8: Male subgroups of interest for pan-cancer analysis

3.4.3 DRUG REPOSITIONING RESULTS

For each unique subgroup of the five subgroups for each gender type, the unique gene patterns of each cancer type for these subgroups were used to create the biomedical network of a given subgroup. For example, the first unique subgroup is females whose age is greater than 65 (SG1-F). This subgroup appeared in 3 types of cancer, which are colon, lung, and ovary. The DR algorithm used genes unique to each cancer type to create the biomedical network for the first subgroup in each cancer type. This process resulted in 3 lists, each of which represents the candidate drugs for that subgroup in the corresponding cancer type. A drug candidates list was created for each of the 16 subgroups (SOI). After finding the drug candidates for each subgroup within each gender type, the top 25 drugs were selected for each subgroup.

To find the drugs for each unique subgroup in each gender type, the common drugs among the disease types were found within each gender type. For example, in SG1-F, the top 25 drugs were used from each of the resulting three-drug lists. The common drugs among these three lists of 25 drugs were retrieved. The purpose is to find the drugs that can be used for that subgroup regardless of the cancer type. To ensure the drugs are not common only because of the general cancer perturbation, the common drugs were analyzed to find those that are unique to each gender in a given subgroup. The unique drugs for each subgroup gender were retrieved. There were unique drugs for each subgroup except for SG5-F and SG3-M. So, these subgroups were dropped from the analysis. For the remaining three subgroups, the unique drugs for each subgroup between genders were analyzed to find the unique drugs across all the subgroups (The six subgroups, three female and three male subgroups). Finally, one to three drugs were found for the subgroup as unique drugs

to that subgroup. Table 9 shows the drugs for each subgroup. These drugs were analyzed to find the genotypic features for each subgroup across cancer types where that subgroup was significant.

ID	Subgroup		Drug Name	#Genes
SG1-F	Age=greater65 & Gender=female	colon, lung, ovary	Vinblastine	176
			Everolimus	233
			Sunitinib	121
SG1-M	Age=greater65 & Gender=male	brain, colon, kidney	Bosutinib	176
			Varenicline	233
SG2-F	Gender=female & Race=white	brain, lung, ovary	Vorinostat	69
			Levonorgestrel	57
SG2-M	Gender=male & Race=white	kidney, lung, rectum	Etoposide	83
			Mitoxantrone	203
SG3-F	Gender=female & Mutation=KRAS	colon, pancreas	Gemcitabine	371
SG3-M	Gender=male & Mutation=KRAS	colon, pancreas	Quinacrine	327

Table 9: Unique drugs for each subgroup

These drugs were analyzed to find the relevance between the drugs and the diseases to which these drugs were suggested for each subgroup:

1- SG1-F drugs are Vinblastine, Everolimus, and Sunitinib. Vinblastine Sulfate is an FDA-approved drug for Hodgkin lymphoma, kaposi sarcoma, and non-Hodgkin lymphoma. In this analysis, it was recommended to be used in colon, lung, and ovary cancer. Studying this drug in the biomedical literature showed that Vinblastine Sulfate was used as a combined therapy to treat colon cancer cell lines. In vitro study showed a decrease in tumor growth and cancer cell invasion and induction of apoptosis [252]. For lung cancer, Vinblastine Sulfate was used as a combined therapy to improve chemotherapy's effectiveness and reduce drug resistance [253].

Also, its anti-tumor effect on ovarian cancer was studied, and it was found that Vinblastine could induce a growth inhibition in these cell lines [254].

Everolimus is an FDA-approved drug for brain tumors, gastroenteropancreatic neuroendocrine tumors, non-small cell lung cancer, pancreatic cancer, renal cell cancer, and small cell lung cancer. The literature review showed Everolimus, an mTOR inhibitor, was used as a combined therapy to treat CRC patients with a wild-type RAS and mutated PIK3CA [255]. Also, it was used to break the S6K1-IRS-2/PI3K negative feedback loop in BRAF V600E CRC patients, which induced apoptosis [256]. In lung cancer, Everolimus was found to inhibit the proliferation and migration of EGFR-resistant lung cancer cells via the upregulation of PTEN and the downregulation of the miR-4328 signaling pathway [257]. For ovarian cancer, Everolimus has an anti-tumor effect by acting on mTORC1 in ovarian cancer. An in vitro study showed that Everolimus inhibited the proliferation and reduced invasion of the ovarian cancer cells by inhibiting the mTOR pathway [258].

Sunitinib is an FDA-approved drug for gastrointestinal stromal tumors, pancreatic cancer, and renal cell cancer. The biomedical literature stated Sunitinib, a multiple tyrosine kinases inhibitor, induces apoptosis and inhibits cell growth in colon cancer cells [259, 260]. Sunitinib was approved as a treatment for gastrointestinal stromal tumors, and it showed efficacy in phase II trials as a treatment of advanced NSCLC [261, 262].

2- SG1-M drugs are Bosutinib and Varenicline. In this analysis, these drugs showed indications for their ability to be used for the brain, colon, and kidney. Bosutinib is an FDA-approved drug for Leukemia. A literature review was done to find the

potential of Bosutinib as brain, colon, and kidney treatments. Using Bosutinib as combined therapy with Valproic acid (VPA) showed a substantial apoptosis induction in colon cancer cells [263]. A clinical trial showed that bosutinib reduced kidney growth in patients with dominant polycystic kidney disease (ADPKD). It has been studied in brain related disease and as a multi-kinase target showed advantages in alleviating neurodegenerative pathologies [264]. Its effect on kidney and brain cancer needs further study. Varenicline is used as an aid in smoking cessation. In this study, it showed an indication to be beneficial for patients with brain, colon, and kidney.

3- SG2-F drugs are Vorinostat and Levonorgestrel, which recommended as highly contrast subgroups in brain, lung, and ovary cancer. Vorinostat is an FDA-approved drug for Non-Hodgkin Lymphoma. it is a drug currently under study for the treatment of cutaneous T cell lymphoma (CTCL). A recent study suggested that vorinostat also possesses some activity against recurrent glioblastoma multiforme, resulting in a median overall survival of 5.7 months (compared to 4 - 4.4 months in earlier studies). Further brain tumor trials are planned using combinations of vorinostat with other drugs [265, 266]. In literature, Vorinostat was suggested as a combined therapy for human brain cancers [267]. So, it is already presenting promising results for brain cancer. In this study, it appeared to have the potential to be used for lung and ovary cancer. In lung cancer, Vorinostat was studied as a treatment for C797S-resistant lung adenocarcinomas when it was used as a combined therapy and showed an improvement in EGFR-TKI sensitivity to EGFR C797S by inducing EGFR-dependent cell death [268]. In a phase I/Ib clinical trial,

Vorinostat showed anti-cancer activities in lung cancer when it was combined with Pembrolizumab [269]. Also, combining Vorinostat with EGFR-TKI can reverse EGFR-TKI resistance in NSCLC [270]. Levonorgestrel is a hormonal medication. It is not approved yet for any cancer treatment, but its role in cancer treatment has been investigated. For ovarian cancer, Levonorgestrel in a phase II clinical trial significantly decreased the proliferation in the ovarian epithelium [271].

4- In SG2-M, Etoposide and Mitoxantrone were the candidate drugs for kidney, lung, and rectum cancer. Etoposide is an FDA-approved drug for small cell lung cancer. Etoposide has anticancer activities as an inhibitor of Topoisomerase II (Topo II) activity in kidney cancer [272]. Etoposide was used as a combined therapy for SCLC patients with sensitive and refractory relapse and showed anticancer activity and was suggested as a second-line treatment option for SCLC patients [273]. In a clinical trial, using Etoposide improved overall survival in patients with extensive-stage small-cell lung cancer (ES-SCLC) [274, 275]. Etoposide was found to have the potential as a treatment for metastatic colorectal cancer when it was used as a combined therapy [276]. Mitoxantrone is an FDA-approved drug for Leukemia, and prostate cancer. There are studies that examined Mitoxantrone potential in other cancer types like lung and colorectal cancer. In NSCLC, Mitoxantrone was found to induce cell apoptosis by phosphorylating ROS1 and inhibit its downstream signaling pathway [277]. A recent study showed that using Mitoxantrone as a combined therapy is a promising treatment for CRC patients due to its inhibitory effects on CRC and autophagic cell death [278].

5- In SG3-F, Gemcitabine was the candidate drug for colon and pancreas cancer. This drug is an FDA-approved drug for non-small cell lung cancer, Ovarian, and Pancreatic Cancer. Gemcitabine is used to induce apoptosis in CRC patients [279]. In a pancreas clinical trial, Gemcitabine has been considered a first-line chemotherapy treatment [280]. Gemcitabine showed effectiveness in mitigation of some systems associated with advanced pancreas cancer and a modest enhancement in survival [281].
6- In SG3-M, Quinacrine was the candidate drug for colon and pancreas cancer. Quinacrine, FDA-approved the drug as anti-malaria therapy, has shown anti-cancer activities. In CRC, the increase of Quinacrine concentration increased the apoptosis induction in CRC cell lines [282].

Analyzing the drug targets in these subgroups across the cancer types for each subgroup showed that there are common targets for these drugs in different types of cancer where that subgroup was significance. These common drug targets are either the abnormally expressed genes in a given subgroup or genes that have direct interaction with the abnormally expressed genes. On average, there were 177 genes in SG1-F, 136 genes in SG1-M, 42 genes in SG2-F, 95 genes in SG2-M, 371 genes in SG3-F, and 327 genes in SG3-M.

3.5 DISCUSSION

Cancer is a group of diseases characterized by the abnormal growth of cells. The existence of common characteristics among these different types of cancer indicates the importance of pan-cancer analysis to study the inter heterogeneity in cancer. This part of the study aims to find homogeneous subpopulations that share genotypic and phenotypic characteristics regardless of the cancer type. Finding homogeneous subgroups will help identify drugs that can treat patients across cancer types and reduce the healthcare cost associated with developing different drugs for different subgroups of patients in each cancer type. Still, developing de novo drugs is expensive and time-consuming. In this pan-cancer analysis, the stratification and drug repositioning algorithm was applied to identify homogeneous subgroups of patients across cancer types and identify candidate drugs for each subgroup.

From TCGA, the genotypic and phenotypic data for 3983 patients across 16 cancer types in 11 organs were obtained. The patient stratification and drug repositioning framework was implemented to find the subgroups. Then, using the genotypic features of these homogeneous subgroups, drugs were recommended for each subgroup based on the genotypic features and their connection to other biomedical entities like pathways, diseases, GO terms, etc. After filtering the resulting subgroup, gender was the phenotypic feature selected to explore the subgroups where gender was a significant phenotypic variable. To have a comparable result, the subgroups that appeared in both genders regardless the cancer type were selected to be analyzed. This resulted in 50 subgroups (25 for each gender). Because the goal is to find the common mechanism that should be targeted across cancer types, the final set of subgroups was the one that appeared in both

genders, and was significant in more than one cancer type. This filtering resulted in five subgroups for each gender to be analyzed.

The drugs were recommended for these subgroups based on the unique, abnormal genes for each disease type in each subgroup. The common drugs between the disease types of each subgroup were analyzed to find the drugs that are suitable for a given subgroup across cancer types. The unique drugs for each subgroup gender were retrieved. The analysis of the recommended drugs for each subgroup showed that these drugs could be repurposed for more than one cancer type. Also, the Drug-Gene interactions for these drugs were analyzed to find the shared genotypic mechanism between different types of cancer for each subgroup. The average number of abnormally expressed genes that are common drug targets across cancer types was 186 genes. Most of the drugs are FDA-approved for different types of cancer, and there is ongoing research to study their potential on other cancer types, including the types to which they were recommended in this study. Further analysis is needed in the context of wet lab experiments and clinical trials to validate these results before recommending them for clinical use.

Chapter 4
Conclusion and Future Work

4.1 CONCLUSION

This dissertation aims to stratify a disease population based on the genotypic and phenotypic features into homogeneous subpopulations and reposition drugs for each subgroup. To achieve that, an explainable artificial intelligence (XAI) framework was developed and implemented. The design, implementation, and validation results demonstrate the potential of the XAI framework to stratify patients into druggable subgroups and to reposition drugs for these subgroups based on their genotypic features after considering their relation to other biomedical entities such as protein-protein interaction, pathways, biological process, cellular component, molecular function, and drugs. Stratifying patients using genotypic and phenotypic features, the explainability of the results, and the ability to provide candidate drugs for each subgroup to overcome the limitation of developing drugs for a small portion of disease represent a promising contribution to improve patient care in our healthcare system by implementing personalized medicine.

In Chapter Two, the development of the patient stratification and the drug repositioning framework was introduced. The stratification process consists of a three-layer system where data mining was used to find homogeneous phenotypic and genotypic features within a heterogeneous disease population. After mapping the genotypic feature into a heterogeneous knowledge base to find druggable subgroups, a network processing framework was developed to consider the relations between the genes and other biomedical entities. Using the homogeneous genotypic features of each subgroup, a drug repositioning framework was developed to recommend drugs for each subgroup. The recommended

drugs were ranked using an innovative drug scoring function that considers various factors to assess drugs potential and relevance to each subgroup.

In Chapter Three, the implementation of patient stratification and drug repositioning was presented. The implementation was performed on different datasets from the TCGA database. First, this method was implemented on colorectal cancer data. The subgroups with MS status were analyzed, and drug recommendations were validated using biomedical literature and computationally using randomized analysis. Then, breast cancer data was used to find homogeneous subgroups within TNBC and recommending drugs for each subgroup. Based on the targeted mechanism by the candidate drugs for each subgroup, age was found to be an important factor in deciding the treatment. Ferroptosis was recommended to be the targeted mechanism in the older population, and antioxidant was recommended to be targeted in the subgroup younger than 50 years old. The third part of this chapter is the pan-cancer analysis. The data for 16 types of cancer in 11 organs were used to find homogeneous cancer subgroups out of this cancer population. The subgroups were filtered, and the gender subgroups were selected for further analysis. The literature review was used to find the potential of these drugs across cancer types. Many of the drugs were FDA-approved drugs for cancer, and they were subjects for a wet lab experiment or clinical trials due to their potential in treating different types of cancer. Some candidate drugs in this analysis were not found as a part of any clinical trial. These drugs may be the novel candidates that should be studied further to evaluate their possibility as a treatment for the recommended cancer types.

4.2 LIMITATIONS

This is a data driven approach and the availability of the data represents a crucial need to do this kind of analysis. One of the limitations of this study is the availability of open access datasets. This could be a limitation for any other data driven approach, but what makes it more challenging in this study is the requirement to have both phenotypic and genotypic data for a large number of patients because this is a data mining-based analysis.

The other significant challenge that this study faced in all the implementations is the validation in a wet lab setting. To overcome this limitation, a literature review was used to find the biomedical merit for the results. Still, wet lab experimentation and clinical trial validation is needed before implementing any of our findings on patients.

4.3 CONTRIBUTION TO INFORMATICS AND CANCER RESEARCH

This work aims to find homogeneous subgroups of patients within a disease population to implement precision medicine in our healthcare system to improve patient survival and reduce treatment costs. This was addressed in this study as follows:

1- *Providing a multi-dimensional patient stratification method:* The developed method uses heterogeneous data types to represent the biological system. The genotypic data, phenotypic data, and biomedical entities were used to stratify patients and find groups of patients that share phenotypic and genotypic similarities. At the same time, they have significant contrast from the rest of the disease population. Finding these subgroups will improve drug efficiency, where the drugs will be used only for the patients who can benefit from them because the drug targets the common mechanism among the patients in that subgroup.

2- *Finding the candidate drugs for homogeneous subgroups*: This study developed a drug repositioning algorithm in which different biomedical entities were taken into consideration in order to find the drugs that can be used to target the abnormal mechanism for each subgroup of patients using their genotypic characteristics and the relation between the genes and other biomedical entities like the pathways, molecular functions, biological processes, cellular components, tissues, and diseases.

3- *Presenting explainable results for the medical practitioners*: The explainability of this method demonstrates the ability to explain the reasoning behind the selection of the subgroups and drugs. Its explainability offers the possibility to understand the mechanism of action to address drug resistance and to find combined therapy. This represents an important factor in ensuring the implementation of the method in the clinical setting because applying black-box methods is challenging due to the lack of the explainability of the results. This explainability is crucial for medical practitioners to decide if a patient or a group of patients can be treated with any recommended drug when referring to a computational method.

4- *Reducing the cost, risk, and time associated with de novo drugs development*: Developing de novo drugs is a time-consuming, high-cost, high-risk process with a low FDA approval rate. This makes it challenging to recommend developing a new drug for a subgroup of patients within a disease population. This study represents a way of solving this problem by finding new indications for drugs that have already been approved and declared safe for human use. The drug

repositioning algorithm presented in this study represents a promising approach to overcoming the difficulties of developing new drugs.

4.4 FUTURE WORK

The future work will continue to add more perspective to the drug repositioning framework. Drug structure information will be added to the biomedical knowledge based to include the binding information of the drugs that could be useful to reposition drugs for patients with mutated genes. Also, the drug knowledge base will be updated to include recently approved drugs. To increase the scope of this framework implementation, a web-based tool will be developed to ensure easy access to this framework by other researchers in the scientific community. In the next phase of this research, the plan is to validate these promising results at the bench in tumor cell lines in vitro and in vivo. In addition, drug resistance mechanisms and side effect profiles will be further studied in preparation for clinical translation.

BIBLIOGRAPHY

1. Deotarse, P., et al., Drug repositioning: a review. Int. J. Pharma. Res Rev, 2015. 4: p. 51-8.
2. Alaimo, S. and A. Pulvirenti, Network-Based Drug Repositioning: Approaches, Resources, and Research Directions. Methods Mol Biol, 2019. 1903: p. 97-113.
3. Plenge, R.M., E.M. Scolnick, and D. Altshuler, Validating therapeutic targets through human genetics. Nat Rev Drug Discov, 2013. 12(8): p. 581-94.
4. Liu, Z., et al., In silico drug repositioning: what we need to know. Drug Discov Today, 2013. 18(3-4): p. 110-5.
5. Metaphor, S.B. and S.B.D. Nicholas, Time for one-person trials. NATURE, 2015. 520.
6. Park, K., A review of computational drug repurposing. Translational and Clinical Pharmacology, 2019. 27(2): p. 59-63.
7. Readhead, B. and J. Dudley, Translational Bioinformatics Approaches to Drug Development. Adv Wound Care (New Rochelle), 2013. 2(9): p. 470-489.
8. Keserci, S., et al., Research synergy and drug development: Bright stars in neighboring constellations. Heliyon, 2017. 3(11): p. e00442.
9. Xue, H., et al., Review of drug repositioning approaches and resources. International journal of biological sciences, 2018. 14(10): p. 1232.
10. Xu, R. and Q. Wang, Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinformatics, 2013. 14: p. 181.

11. Lotfi Shahreza, M., et al., Heter-LP: A heterogeneous label propagation algorithm and its application in drug repositioning. J Biomed Inform, 2017. 68: p. 167-183.
12. Hu, G. and P. Agarwal, Human disease-drug network based on genomic expression profiles. PLoS One, 2009. 4(8): p. e6536.
13. Xu, R. and Q. Wang, PhenoPredict: A disease phenome-wide drug repositioning approach towards schizophrenia drug discovery. J Biomed Inform, 2015. 56: p. 348-55.
14. Campillos, M., et al., Drug target identification using side-effect similarity. Science, 2008. 321(5886): p. 263-266.
15. Xu, R. and Q. Wang, A genomics-based systems approach towards drug repositioning for rheumatoid arthritis. BMC Genomics, 2016. 17 Suppl 7(Suppl 7): p. 518.
16. Lamb, J., et al., The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. science, 2006. 313(5795): p. 1929-1935.
17. Lamb, J., The Connectivity Map: a new tool for biomedical research. Nature reviews cancer, 2007. 7(1): p. 54.
18. Liu, H., et al., Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. BMC Bioinformatics, 2016. 17(Suppl 17): p. 539.
19. Wang, Y., et al., Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. PloS one, 2013. 8(11): p. e78518.

20. Tian, Z., et al., Computational drug repositioning using meta-path-based semantic network analysis. BMC Syst Biol, 2018. 12(Suppl 9): p. 134.
21. Lee, B.K., et al., DeSigN: connecting gene expression with therapeutics for drug repurposing and development. BMC Genomics, 2017. 18(Suppl 1): p. 934.
22. Cheng, F., et al., A network-based drug repositioning infrastructure for precision cancer medicine through targeting significantly mutated genes in the human cancer genomes. J Am Med Inform Assoc, 2016. 23(4): p. 681-91.
23. Liu, D., et al., Exploratory Data Mining for Subgroup Cohort Discoveries and Prioritization. IEEE journal of biomedical and health informatics, 2019.
24. Himmelstein, D.S., et al., Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife, 2017. 6: p. e26726.
25. Schneider, L., et al., ClinOmicsTrailbc: a visual analytics tool for breast cancer treatment stratification. Bioinformatics, 2019.
26. Chen, Y. and R. Xu, Drug repurposing for glioblastoma based on molecular subtypes. J Biomed Inform, 2016. 64: p. 131-138.
27. Turanli, B., et al., Multi-Omic Data Interpretation to Repurpose Subtype Specific Drug Candidates for Breast Cancer. Frontiers in genetics, 2019. 10: p. 420.
28. Zhou, C., et al., Statin use and its potential therapeutic role in esophageal cancer: a systematic review and meta-analysis. Cancer Manag Res, 2019. 11: p. 5655-5663.
29. Gouravan, S., et al., Preclinical Evaluation of Vemurafenib as Therapy for BRAF(V600E) Mutated Sarcomas. Int J Mol Sci, 2018. 19(4).

30. Simon, L., et al., Chemogenomic Landscape of RUNX1-mutated AML Reveals Importance of RUNX1 Allele Dosage in Genetics and Glucocorticoid Sensitivity. Clin Cancer Res, 2017. 23(22): p. 6969-6981.

31. Yoshida, G.J., Emerging roles of Myc in stem cell biology and novel tumor therapies. J Exp Clin Cancer Res, 2018. 37(1): p. 173.

32. Nepal, C., et al., Genomic perturbations reveal distinct regulatory networks in intrahepatic cholangiocarcinoma. Hepatology, 2018. 68(3): p. 949-963.

33. Lind, A.P. and P.C. Anderson, Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. PLoS One, 2019. 14(7): p. e0219774.

34. GLIGORIJEVIĆ, V., N. Malod-Dognin, and N. PRŽULJ. Patient-specific data fusion for cancer stratification and personalised treatment. in Biocomputing 2016: Proceedings of the Pacific Symposium. 2016. World Scientific.

35. Holzinger, A., A. Carrington, and H. Müller, Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. Kunstliche Intell (Oldenbourg), 2020. 34(2): p. 193-198.

36. Hagras, H., Toward human-understandable, explainable AI. Computer, 2018. 51(9): p. 28-36.

37. Slomka, P.J., et al., Application and Translation of Artificial Intelligence to Cardiovascular Imaging in Nuclear Medicine and Noncontrast CT. Semin Nucl Med, 2020. 50(4): p. 357-366.

38. Harfouche, A.L., et al., Accelerating Climate Resilient Plant Breeding by Applying Next-Generation Artificial Intelligence. Trends Biotechnol, 2019. 37(11): p. 1217-1235.

39. Holzinger, A., et al., Interactive machine learning: experimental evidence for the human in the algorithmic loop. Applied Intelligence, 2019. 49(7): p. 2401-2414.

40. Wallace, B.C., et al. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. in Proceedings of the 2nd ACM SIGHIT international health informatics symposium. 2012.

41. Teso, S. and K. Kersting. Explanatory interactive machine learning. in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019.

42. Doshi-Velez, F. and B. Kim, Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

43. Al-Taie, Z., et al., Explainable artificial intelligence in high-throughput drug repositioning for subgroup stratifications with interventionable potential. J Biomed Inform, 2021. 118: p. 103792.

44. Dong, G. and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. in Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. 1999.

45. Pudil, P., J. Novovičová, and J. Kittler, Floating search methods in feature selection. Pattern recognition letters, 1994. 15(11): p. 1119-1125.

46. Agarwal, R. and R. Srikant. Fast algorithms for mining association rules. in Proc. of the 20th VLDB Conference. 1994.

47. Jain, A., K. Nandakumar, and A. Ross, Score normalization in multimodal biometric systems. Pattern recognition, 2005. 38(12): p. 2270-2285.

48. Wu, C., et al., Computational drug repositioning through heterogeneous network clustering. BMC Syst Biol, 2013. 7 Suppl 5: p. S6.

49. Yu, L., et al., Conserved Disease Modules Extracted From Multilayer Heterogeneous Disease and Gene Networks for Understanding Disease Mechanisms and Predicting Disease Treatments. Front Genet, 2018. 9: p. 745.

50. Qi, X., et al., The Performance of Gene Expression Signature-Guided Drug-Disease Association in Different Categories of Drugs and Diseases. Molecules, 2020. 25(12).

51. Lee, T. and Y. Yoon, Drug repositioning using drug-disease vectors based on an integrated network. BMC Bioinformatics, 2018. 19(1): p. 446.

52. Iwata, M., et al., Pathway-Based Drug Repositioning for Cancers: Computational Prediction and Experimental Validation. J Med Chem, 2018. 61(21): p. 9583-9595.

53. Wu, Z., Y. Wang, and L. Chen, Drug repositioning framework by incorporating functional information. IET Syst Biol, 2013. 7(5): p. 188-94.

54. Kissa, M., G. Tsatsaronis, and M. Schroeder, Prediction of drug gene associations via ontological profile similarity with application to drug repositioning. Methods, 2015. 74: p. 71-82.

55. Zheng, Y., et al., Old drug repositioning and new drug discovery through similarity learning from drug-target joint feature spaces. BMC Bioinformatics, 2019. 20(Suppl 23): p. 605.

56. Taguchi, Y. and T. Turki, Universal Nature of Drug Treatment Responses in Drug-Tissue-Wide Model-Animal Experiments Using Tensor Decomposition-Based Unsupervised Feature Extraction. Front Genet, 2020. 11: p. 695.

57. Liu, Z., et al., Similarity-based prediction for Anatomical Therapeutic Chemical classification of drugs by integrating multiple data sources. Bioinformatics, 2015. 31(11): p. 1788-95.

58. Hameed, P.N., et al., A two-tiered unsupervised clustering approach for drug repositioning through heterogeneous data integration. BMC Bioinformatics, 2018. 19(1): p. 129.

59. Sun, Y., et al., A physarum-inspired prize-collecting steiner tree approach to identify subnetworks for drug repositioning. BMC Syst Biol, 2016. 10(Suppl 5): p. 128.

60. Markowitz, S.D. and M.M. Bertagnolli, Molecular basis of colorectal cancer. New England Journal of Medicine, 2009. 361(25): p. 2449-2460.

61. Vogelstein, B., et al., Genetic alterations during colorectal-tumor development. New England Journal of Medicine, 1988. 319(9): p. 525-532.

62. Hasan, S., et al., Microsatellite instability (MSI) as an independent predictor of pathologic complete response (PCR) in locally advanced rectal cancer: a National Cancer Database (NCDB) Analysis. Annals of surgery, 2020. 271(4): p. 716-723.

63. André, T., et al., Adjuvant fluorouracil, leucovorin, and oxaliplatin in stage II to III colon cancer: updated 10-year survival and outcomes according to BRAF mutation and mismatch repair status of the MOSAIC study. Journal of Clinical Oncology, 2015. 33(35): p. 4176-4187.

64. Le, D.T., et al., PD-1 blockade in tumors with mismatch-repair deficiency. New England Journal of Medicine, 2015. 372(26): p. 2509-2520.

65. Church, K. and W. Gale, Inverse document frequency (idf): A measure of deviations from poisson, in Natural language processing using very large corpora. 1999, Springer. p. 283-295.

66. Al-Taie, Z., et al., Drug Repositioning and Subgroup Discovery for Precision Medicine Implementation in Triple Negative Breast Cancer. Cancers (Basel), 2021. 13(24).

67. Al-Taie, Z., et al., Explainable artificial intelligence in high-throughput drug repositioning for subgroup stratifications with interventionable potential. Journal of Biomedical Informatics, 2021. 118: p. 103792.

68. Murcia, O., et al., Colorectal cancer molecular classification using BRAF, KRAS, microsatellite instability and CIMP status: Prognostic implications and response to chemotherapy. PLoS One, 2018. 13(9): p. e0203051.

69. Le, D.T., et al., PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med, 2015. 372(26): p. 2509-20.

70. Loupakis, F., et al., Primary tumor location as a prognostic factor in metastatic colorectal cancer. J Natl Cancer Inst, 2015. 107(3).

71. Nyamundanda, G., E. Fontana, and A. Sadanandam, Is the tumour microenvironment a critical prognostic factor in early-stage colorectal cancer? Ann Oncol, 2019. 30(10): p. 1538-1540.

72. Yang, Y., et al., Gender differences in colorectal cancer survival: A meta-analysis. Int J Cancer, 2017. 141(10): p. 1942-1949.

73. Dienstmann, R., et al., Relative contribution of clinicopathological variables, genomic markers, transcriptomic subtyping and microenvironment features for outcome prediction in stage II/III colorectal cancer. Ann Oncol, 2019. 30(10): p. 1622-1629.

74. Kishore, C., S. Sundaram, and D. Karunagaran, Vitamin K3 (menadione) suppresses epithelial-mesenchymal-transition and Wnt signaling pathway in human colorectal cancer cells. Chem Biol Interact, 2019. 309: p. 108725.

75. Hegazy, M.F., et al., Vitamin K(3) thio-derivative: a novel specific apoptotic inducer in the doxorubicin-sensitive and -resistant cancer cells. Invest New Drugs, 2020. 38(3): p. 650-661.

76. Nakamura, Y. and T. Yamaguchi, Stereoselective metabolism of 2-phenylpropionic acid in rat. I. In vitro studies on the stereoselective isomerization and glucuronidation of 2-phenylpropionic acid. Drug Metab Dispos, 1987. 15(4): p. 529-34.

77. Du, F., et al., SOX13 promotes colorectal cancer metastasis by transactivating SNAI2 and c-MET. Oncogene, 2020. 39(17): p. 3522-3540.

78. Graves-Deal, R., et al., Broad-spectrum receptor tyrosine kinase inhibitors overcome de novo and acquired modes of resistance to EGFR-targeted therapies in colorectal cancer. Oncotarget, 2019. 10(13): p. 1320-1333.

79. Cuneo, K.C., et al., Enhancing the Radiation Response in KRAS Mutant Colorectal Cancers Using the c-Met Inhibitor Crizotinib. Transl Oncol, 2019. 12(2): p. 209-216.

80. Sordet, O., et al., Apoptosis induced by topoisomerase inhibitors. Curr Med Chem Anticancer Agents, 2003. 3(4): p. 271-90.

81. Dehshahri, A., et al., Topoisomerase inhibitors: Pharmacology and emerging nanoscale delivery systems. Pharmacol Res, 2020. 151: p. 104551.

82. Jacob, S., et al., The role of the DNA mismatch repair system in the cytotoxicity of the topoisomerase inhibitors camptothecin and etoposide to human colorectal cancer cells. Cancer Res, 2001. 61(17): p. 6555-62.

83. Stintzing, S. and H.J. Lenz, Protein kinase inhibitors in metastatic colorectal cancer. Let's pick patients, tumors, and kinase inhibitors to piece the puzzle together! Expert Opin Pharmacother, 2013. 14(16): p. 2203-20.

84. Nygård, S.B., et al., Underpinning the repurposing of anthracyclines towards colorectal cancer: assessment of topoisomerase II alpha gene copy number alterations in colorectal cancer. Scand J Gastroenterol, 2013. 48(12): p. 1436-43.

85. Tarpgaard, L.S., et al., A phase II study of Epirubicin in oxaliplatin-resistant patients with metastatic colorectal cancer and TOP2A gene amplification. BMC cancer, 2016. 16(1): p. 1-5.

86. Lai, J.I., et al., Clinical Perspective of FDA Approved Drugs With P-Glycoprotein Inhibition Activities for Potential Cancer Therapeutics. Front Oncol, 2020. 10: p. 561936.

87. Chen, E.Y., et al., Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics, 2013. 14: p. 128.

88. Kuleshov, M.V., et al., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res, 2016. 44(W1): p. W90-7.

89. Fan, Q. and B. Liu, Identification of the anticancer effects of a novel proteasome inhibitor, ixazomib, on colorectal cancer using a combined method of microarray and bioinformatics analysis. Onco Targets Ther, 2017. 10: p. 3591-3606.

90. Ye, L.C., et al., Downregulated long non-coding RNA CLMAT3 promotes the proliferation of colorectal cancer cells by targeting regulators of the cell cycle pathway. Oncotarget, 2016. 7(37): p. 58931-58938.

91. Tan, J., et al., Pharmacologic modulation of glycogen synthase kinase-3beta promotes p53-dependent apoptosis through a direct Bax-mediated mitochondrial pathway in colorectal cancer cells. Cancer Res, 2005. 65(19): p. 9012-20.

92. Richardson, C., et al., Small-molecule CB002 restores p53 pathway signaling and represses colorectal cancer cell growth. Cell Cycle, 2017. 16(18): p. 1719-1725.

93. Attoub, S., et al., The c-kit tyrosine kinase inhibitor STI571 for colorectal cancer therapy. Cancer Res, 2002. 62(17): p. 4879-83.

94. Quinn, B.J., et al., Repositioning metformin for cancer prevention and treatment. Trends in Endocrinology and Metabolism, 2013. 24(9): p. 469-480.

95. Saini, N. and X. Yang, Metformin as an anti-cancer agent: actions and mechanisms targeting cancer stem cells. Acta Biochim Biophys Sin (Shanghai), 2018. 50(2): p. 133-143.

96. Wojciechowska, J., et al., Diabetes and Cancer: a Review of Current Knowledge. Exp Clin Endocrinol Diabetes, 2016. 124(5): p. 263-75.

97. Jones, G.R. and M.P. Molloy, Metformin, Microbiome and Protection Against Colorectal Cancer. Dig Dis Sci, 2020.

98. Gadducci, A., et al., Metformin use and gynecological cancers: A novel treatment option emerging from drug repositioning. Crit Rev Oncol Hematol, 2016. 105: p. 73-83.

99. De Pauw, I., et al., Overcoming Intrinsic and Acquired Cetuximab Resistance in RAS Wild-Type Colorectal Cancer: An In Vitro Study on the Expression of HER Receptors and the Potential of Afatinib. Cancers (Basel), 2019. 11(1).

100. Yang, M., et al., Afatinib treatment for her-2 amplified metastatic colorectal cancer based on patient-derived xenograft models and next generation sequencing. Cancer Biol Ther, 2019. 20(4): p. 391-396.

101. Dunn, E.F., et al., Dasatinib sensitizes KRAS mutant colorectal tumors to cetuximab. Oncogene, 2011. 30(5): p. 561-74.

102. Rao, G., et al., Dasatinib sensitises KRAS-mutant cancer cells to mitogen-activated protein kinase kinase inhibitor via inhibition of TAZ activity. Eur J Cancer, 2018. 99: p. 37-48.

103. Williams, C.B., et al., A metastatic colon adenocarcinoma harboring BRAF V600E has a durable major response to dabrafenib/trametinib and chemotherapy. Onco Targets Ther, 2015. 8: p. 3561-4.

104. Leung, S.W., et al., An Integrated Bioinformatics Analysis Repurposes an Antihelminthic Drug Niclosamide for Treating HMGA2-Overexpressing Human Colorectal Cancer. Cancers (Basel), 2019. 11(10).

105. Isacoff, W.H. and K. Borud, Chemotherapy for the treatment of patients with metastatic colorectal cancer: an overview. World J Surg, 1997. 21(7): p. 748-62.

106. Pawlak, A., et al., Long-lasting reduction in clonogenic potential of colorectal cancer cells by sequential treatments with 5-azanucleosides and topoisomerase inhibitors. BMC Cancer, 2016. 16(1): p. 893.

107. Kunnumakkara, A.B., et al., Cancer drug development: The missing links. Experimental Biology and Medicine, 2019. 244(8): p. 663-689.

108. Hicks, S.C., et al., Smooth quantile normalization. Biostatistics, 2018. 19(2): p. 185-198.

109. Turashvili, G. and E. Brogi, Tumor Heterogeneity in Breast Cancer. Front Med (Lausanne), 2017. 4: p. 227.

110. Tian, K., et al., p53 modeling as a route to mesothelioma patients stratification and novel therapeutic identification. Journal of translational medicine, 2018. 16(1): p. 1-15.

111. Louhimo, R., et al., Data integration to prioritize drugs using genomics and curated data. BioData Min, 2016. 9: p. 21.

112. Jin, G., et al., A novel method of transcriptional response analysis to facilitate drug repositioning for cancer therapy. Cancer research, 2012. 72(1): p. 33-44.

113. Cava, C., G. Bertoli, and I. Castiglioni, In silico identification of drug target pathways in breast cancer subtypes using pathway cross-talk inhibition. Journal of translational medicine, 2018. 16(1): p. 1-17.

114. Carrella, D., et al., Computational drugs repositioning identifies inhibitors of oncogenic PI3K/AKT/P70S6K-dependent pathways among FDA-approved compounds. Oncotarget, 2016. 7(37): p. 58743.

115. Wong, H.S., et al., Integrative bioinformatic analyses of an oncogenomic profile reveal the biology of endometrial cancer and guide drug discovery. Oncotarget, 2016. 7(5): p. 5909-23.

116. Peyvandipour, A., et al., A novel computational approach for drug repurposing using systems biology. Bioinformatics, 2018. 34(16): p. 2817-2825.

117. Wang, J., et al., Identification of associations between small molecule drugs and miRNAs based on functional similarity. Oncotarget, 2016. 7(25): p. 38658.

118. Yang, J., et al., Drug–disease association and drug-repositioning predictions in complex diseases using causal inference–probabilistic matrix factorization. Journal of chemical information and modeling, 2014. 54(9): p. 2562-2569.

119. Fahimian, G., et al., RepCOOL: computational drug repositioning via integrating heterogeneous biological networks. Journal of Translational Medicine, 2020. 18(1): p. 1-10.

120. Klahan, S., et al., Identification of genes and pathways related to lymphovascular invasion in breast cancer patients: A bioinformatics analysis of gene expression profiles. Tumour Biol, 2017. 39(6): p. 1010428317705573.

121. Karuppasamy, R., et al., An Integrative drug repurposing pipeline: switching viral drugs to breast cancer. Journal of cellular biochemistry, 2017. 118(6): p. 1412-1422.

122. Rymbai, E., et al., Ropinirole, a potential drug for systematic repositioning based on side effect profile for management and treatment of Breast Cancer. Medical Hypotheses, 2020. 144: p. 110156.

123. Zhao, H., et al., Novel modeling of cancer cell signaling pathways enables systematic drug repositioning for distinct breast cancer metastases. Cancer research, 2013. 73(20): p. 6149-6163.

124. Yu, L., J. Zhao, and L. Gao, Predicting Potential Drugs for Breast Cancer based on miRNA and Tissue Specificity. Int J Biol Sci, 2018. 14(8): p. 971-982.

125. Jadamba, E. and M. Shin, A Systematic Framework for Drug Repositioning from Integrated Omics and Drug Phenotype Profiles Using Pathway-Drug Network. Biomed Res Int, 2016. 2016: p. 7147039.

126. Cava, C., S. Sabetian, and I. Castiglioni, Patient-Specific Network for Personalized Breast Cancer Therapy with Multi-Omics Data. Entropy, 2021. 23(2): p. 225.

127. Zhu, L. and J. Liu, Integration of a prognostic gene module with a drug sensitivity module to identify drugs that could be repurposed for breast cancer therapy. Computers in biology and medicine, 2015. 61: p. 163-171.

128. Johnson, K.S., E.F. Conant, and M.S. Soo, Molecular subtypes of breast cancer: a review for breast radiologists. Journal of Breast Imaging, 2021.

129. Cava, C., et al., Identification of Breast Cancer Subtype-Specific Biomarkers by Integrating Copy Number Alterations and Gene Expression Profiles. Medicina, 2021. 57(3): p. 261.

130. Low, Y.S., et al., Synergistic drug combinations from electronic health records and gene expression. Journal of the American Medical Informatics Association, 2017. 24(3): p. 565-576.

131. Bourdakou, M.M., E.I. Athanasiadis, and G.M. Spyrou, Discovering gene re-ranking efficiency and conserved gene-gene relationships derived from gene co-expression network analysis on breast cancer data. Scientific reports, 2016. 6(1): p. 1-29.

132. Bertoli, G., C. Cava, and I. Castiglioni, The potential of miRNAs for diagnosis, treatment and monitoring of breast cancer. Scandinavian Journal of Clinical and Laboratory Investigation, 2016. 76(sup245): p. S34-S39.

133. Sarkar, J.P., et al., Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers. Computers in Biology and Medicine, 2021. 131: p. 104244.

134. Zhou, S., et al., Systematical analysis of lncRNA–mRNA competing endogenous RNA network in breast cancer subtypes. Breast cancer research and treatment, 2018. 169(2): p. 267-275.

135. Mejía-Pedroza, R.A., J. Espinal-Enríquez, and E. Hernández-Lemus, Pathway-based drug repositioning for breast cancer molecular subtypes. Frontiers in pharmacology, 2018. 9: p. 905.

136. Warchal, S.J., et al., High content phenotypic screening identifies serotonin receptor modulators with selective activity upon breast cancer cell cycle and cytokine signaling pathways. Bioorganic & medicinal chemistry, 2020. 28(1): p. 115209.

137. Metzger-Filho, O., et al., Dissecting the heterogeneity of triple-negative breast cancer. J Clin Oncol, 2012. 30(15): p. 1879-87.

138. Kennecke, H., et al., Metastatic behavior of breast cancer subtypes. J Clin Oncol, 2010. 28(20): p. 3271-7.

139. Kassam, F., et al., Survival outcomes for patients with metastatic triple-negative breast cancer: implications for clinical practice and trial design. Clin Breast Cancer, 2009. 9(1): p. 29-33.

140. Klahan, S., et al., Gene expression profiling combined with functional analysis identify integrin beta1 (ITGB1) as a potential prognosis biomarker in triple negative breast cancer. Pharmacol Res, 2016. 104: p. 31-7.

141. Chang, W.C., et al., The association between single-nucleotide polymorphisms of ORAI1 gene and breast cancer in a Taiwanese population. ScientificWorldJournal, 2012. 2012: p. 916587.

142. Klahan, S., et al., Computational analysis of mRNA expression profiles identifies the ITG family and PIK3R3 as crucial genes for regulating triple negative breast cancer cell migration. Biomed Res Int, 2014. 2014: p. 536591.

143. Gradishar, W.J., et al., Breast Cancer, Version 3.2020, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw, 2020. 18(4): p. 452-478.

144. Brenton, J.D., et al., Molecular classification and molecular forecasting of breast cancer: ready for clinical application? Journal of clinical oncology, 2005. 23(29): p. 7350-7360.

145. Li, X., et al., Triple-negative breast cancer has worse overall survival and cause-specific survival than non-triple-negative breast cancer. Breast cancer research and treatment, 2017. 161(2): p. 279-287.

146. Anders, C.K., et al., The evolution of triple-negative breast cancer: from biology to novel therapeutics. American Society of Clinical Oncology Educational Book, 2016. 36: p. 34-42.

147. Jhan, J.-R. and E.R. Andrechek, Triple-negative breast cancer and the potential for targeted therapy. Pharmacogenomics, 2017. 18(17): p. 1595-1609.

148. Vitali, F., et al., A network-based data integration approach to support drug repurposing and multi-target therapies in triple negative breast cancer. PloS one, 2016. 11(9): p. e0162407.

149. Li, D. and Y. Li, The interaction between ferroptosis and lipid metabolism in cancer. Signal transduction and targeted therapy, 2020. 5(1): p. 1-10.

150. Liu, Z., et al., Systematic Analysis of the Aberrances and Functional Implications of Ferroptosis in Cancer. iScience, 2020. 23(7): p. 101302.

151. Liu, Z., et al., Systematic Pan-Cancer Analysis Reveals the Functional Roles of Ferroptosis Across Cancers. bioRxiv, 2019: p. 765826.

152. Solca, F., et al., Target binding properties and cellular activity of afatinib (BIBW 2992), an irreversible ErbB family blocker. Journal of Pharmacology and Experimental Therapeutics, 2012. 343(2): p. 342-350.

153. Bernsdorf, M., et al., Effect of adding gefitinib to neoadjuvant chemotherapy in estrogen receptor negative early breast cancer in a randomized phase II trial. Breast cancer research and treatment, 2011. 126(2): p. 463-470.

154. Girgert, R., G. Emons, and C. Gründker, 17β-estradiol-induced growth of triple-negative breast cancer cells is prevented by the reduction of GPER expression after treatment with gefitinib. Oncology reports, 2017. 37(2): p. 1212-1218.

155. McLaughlin, R.P., et al., A kinase inhibitor screen identifies a dual cdc7/CDK9 inhibitor to sensitise triple-negative breast cancer to EGFR-targeted therapy. Breast Cancer Research, 2019. 21(1): p. 1-15.

156. Malumbres, M., Cyclins and related kinases in cancer cells. Journal of BU ON.: official journal of the Balkan Union of Oncology, 2007. 12: p. S45-52.

157. Zhou, Y., et al., The crosstalk between autophagy and ferroptosis: what can we learn to target drug resistance in cancer? Cancer biology & medicine, 2019. 16(4): p. 630.

158. Smidova, V., et al., Nanomedicine of tyrosine kinase inhibitors. Theranostics, 2021. 11(4): p. 1546.

159. Irby, R.B. and T.J. Yeatman, Role of Src expression and activation in human cancer. Oncogene, 2000. 19(49): p. 5636-5642.

160. Ma, S., et al., Ferroptosis is induced following siramesine and lapatinib treatment of breast cancer cells. Cell death & disease, 2016. 7(7): p. e2307-e2307.

161. Ma, S., et al., Ferroptosis and autophagy induced cell death occur independently after siramesine and lapatinib treatment in breast cancer cells. PLoS One, 2017. 12(8): p. e0182921.

162. Rusnak, D.W., et al., The effects of the novel, reversible epidermal growth factor receptor/ErbB-2 tyrosine kinase inhibitor, GW2016, on the growth of human normal and tumor-derived cell lines in vitro and in vivo. Molecular cancer therapeutics, 2001. 1(2): p. 85-94.

163. Wood, E.R., et al., A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor

off-rate, and receptor activity in tumor cells. Cancer research, 2004. 64(18): p. 6652-6659.

164. Villalpando-Rodriguez, G.E., et al., Lysosomal destabilizing drug siramesine and the dual tyrosine kinase inhibitor lapatinib induce a synergistic ferroptosis through reduced heme oxygenase-1 (HO-1) levels. Oxidative medicine and cellular longevity, 2019. 2019.

165. Li, Z., et al., Targeting ferroptosis in breast cancer. Biomarker Research, 2020. 8(1): p. 1-27.

166. Viswanathan, V.S., et al., Dependency of a therapy-resistant state of cancer cells on a lipid peroxidase pathway. Nature, 2017. 547(7664): p. 453-457.

167. Li, B., et al., Emerging mechanisms and applications of ferroptosis in the treatment of resistant cancers. Biomedicine & Pharmacotherapy, 2020. 130: p. 110710.

168. Qi, X.-F., et al., Involvement of oxidative stress in simvastatin-induced apoptosis of murine CT26 colon carcinoma cells. Toxicology letters, 2010. 199(3): p. 277-287.

169. Dixon, S.J. and B.R. Stockwell, The hallmarks of ferroptosis. Annual Review of Cancer Biology, 2019. 3: p. 35-54.

170. Wu, Y., et al., Ferroptosis in cancer treatment: another way to Rome. Frontiers in Oncology, 2020. 10.

171. Liu, J., et al., Autophagy-dependent ferroptosis: machinery and regulation. Cell chemical biology, 2020. 27(4): p. 420-435.

172. Chen, J.J. and L. Galluzzi, Fighting resilient cancers with iron. Trends in cell biology, 2018. 28(2): p. 77-78.

173. Lin, X., et al., The mechanism of ferroptosis and applications in tumor treatment. Frontiers in Pharmacology, 2020. 11: p. 1061.

174. Lai, Y., et al., Cell death-related molecules and biomarkers for renal cell carcinoma targeted therapy. Cancer cell international, 2019. 19(1): p. 1-15.

175. Mou, Y., et al., Ferroptosis, a new form of cell death: opportunities and challenges in cancer. Journal of Hematology & Oncology, 2019. 12(1): p. 1-16.

176. Santoni, M., et al., Different effects of sunitinib, sorafenib, and pazopanib on inducing cancer cell death: The role of autophagy. 2013, American Society of Clinical Oncology.

177. Fulda, S., Repurposing anticancer drugs for targeting necroptosis. Cell Cycle, 2018. 17(7): p. 829-832.

178. Mou, Y., et al., The Landscape of Iron Metabolism-Related and Methylated Genes in the Prognosis Prediction of Clear Cell Renal Cell Carcinoma. Frontiers in Oncology, 2020. 10.

179. Yang, W.S., et al., Peroxidation of polyunsaturated fatty acids by lipoxygenases drives ferroptosis. Proceedings of the National Academy of Sciences, 2016. 113(34): p. E4966-E4975.

180. Stamenkovic, A., G.N. Pierce, and A. Ravandi, Phospholipid oxidation products in ferroptotic myocardial cell death. American Journal of Physiology-Heart and Circulatory Physiology, 2019. 317(1): p. H156-H163.

181. Jiang, M., et al., Targeting ferroptosis for cancer therapy: exploring novel strategies from its mechanisms and role in cancers. Translational Lung Cancer Research, 2020. 9(4): p. 1569.

182. Lu, J., et al., Extracellular vesicles from endothelial progenitor cells prevent steroid-induced osteoporosis by suppressing the ferroptotic pathway in mouse osteoblasts based on bioinformatics evidence. Scientific reports, 2019. 9(1): p. 1-18.

183. Sanaei, M., F. Kavoosi, and O. Mansoori, Effect of valproic acid in comparison with vorinostat on cell growth inhibition and apoptosis induction in the human colon cancer SW48 cells in vitro. Experimental Oncology, 2018.

184. Miyamoto, K., et al., xCT inhibition increases sensitivity to vorinostat in a ROS-dependent manner. Cancers, 2020. 12(4): p. 827.

185. Yang, H., et al., Pharmacotranscriptomic Analysis Reveals Novel Drugs and Gene Networks Regulating Ferroptosis in Cancer. Cancers, 2020. 12(11): p. 3273.

186. Panieri, E. and L. Saso, Potential applications of NRF2 inhibitors in cancer therapy. Oxidative medicine and cellular longevity, 2019. 2019.

187. Mishima, E., et al., Drugs repurposed as antiferroptosis agents suppress organ damage, including AKI, by functioning as lipid peroxyl radical scavengers. Journal of the American Society of Nephrology, 2020. 31(2): p. 280-296.

188. Pan, Y., et al., Lipid peroxidation aggravates anti-tuberculosis drug-induced liver injury: Evidence of ferroptosis induction. Biochemical and Biophysical Research Communications, 2020. 533(4): p. 1512-1518.

189. Tsvetkova, D., et al., Antioxidant activity of galantamine and some of its derivatives. Current medicinal chemistry, 2013. 20(36): p. 4595-4608.

190. Nie, F., et al., Apoptotic effect of tannic acid on fatty acid synthase over-expressed human breast cancer cells. Tumor Biology, 2016. 37(2): p. 2137-2143.

191. Zhou, L., et al., FASN, ErbB2-mediated glycolysis is required for breast cancer cell migration. Oncology reports, 2016. 35(5): p. 2715-2722.

192. Chen, T., et al., Fatty acid synthase affects expression of ErbB receptors in epithelial to mesenchymal transition of breast cancer cells and invasive ductal carcinoma. Oncology letters, 2017. 14(5): p. 5934-5946.

193. Farhat, D., et al., Lipoic acid-induced oxidative stress abrogates IGF-1R maturation by inhibiting the CREB/furin axis in breast cancer cell lines. Oncogene, 2020. 39(17): p. 3604-3610.

194. Farhat, D., et al., Lipoic acid decreases breast cancer cell proliferation by inhibiting IGF-1R via furin downregulation. British journal of cancer, 2020. 122(6): p. 885-894.

195. Choi, H.S., et al., Synergistic Tumoricidal Effects of Alpha-Lipoic Acid and Radiotherapy on Human Breast Cancer Cells Via HMGB1. Cancer Research and Treatment, 2020.

196. Kuleshov, M.V., et al., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research, 2016. 44(W1): p. W90-W97.

197. Xie, Z., et al., Gene set knowledge discovery with Enrichr. Current protocols, 2021. 1(3): p. e90.

198. Hanahan, D. and R.A. Weinberg, Hallmarks of cancer: the next generation. cell, 2011. 144(5): p. 646-674.

199. Martinez-Outschoorn, U.E., et al., Cancer metabolism: a therapeutic perspective. Nature reviews Clinical oncology, 2017. 14(1): p. 11.

200. Tennant, D.A., R.V. Durán, and E. Gottlieb, Targeting metabolic transformation for cancer therapy. Nature reviews cancer, 2010. 10(4): p. 267-277.

201. Schulze, A. and A.L. Harris, How cancer metabolism is tuned for proliferation and vulnerable to disruption. Nature, 2012. 491(7424): p. 364-373.

202. Long, J.-P., X.-N. Li, and F. Zhang, Targeting metabolism in breast cancer: How far we can go? World journal of clinical oncology, 2016. 7(1): p. 122.

203. Wang, L., S. Zhang, and X. Wang, The metabolic mechanisms of breast cancer metastasis. Frontiers in Oncology, 2021. 10: p. 2942.

204. Gong, Y., et al., Metabolic-pathway-based subtyping of triple-negative breast cancer reveals potential therapeutic targets. Cell Metabolism, 2021. 33(1): p. 51-64. e9.

205. Lanning, N.J., et al., Metabolic profiling of triple-negative breast cancer cells reveals metabolic vulnerabilities. Cancer & metabolism, 2017. 5(1): p. 1-14.

206. Bevinakoppamath, S., et al., Chemopreventive and anticancer property of selenoproteins in obese breast cancer. Frontiers in Pharmacology, 2021. 12.

207. Wang, H., et al., Recognition of Immune Microenvironment Landscape and Immune-Related Prognostic Genes in Breast Cancer. BioMed Research International, 2020. 2020.

208. Sousa, B., et al., P-cadherin induces anoikis-resistance of matrix-detached breast cancer cells by promoting pentose phosphate pathway and decreasing oxidative stress. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 2020. 1866(12): p. 165964.

209. Schömel, N., et al., UGCG overexpression leads to increased glycolysis and increased oxidative phosphorylation of breast cancer cells. Scientific reports, 2020. 10(1): p. 1-13.

210. Helm, J.S. and R.A. Rudel, Adverse outcome pathways for ionizing radiation and breast cancer involve direct and indirect DNA damage, oxidative stress, inflammation, genomic instability, and interaction with hormonal regulation of the breast. Archives of toxicology, 2020. 94: p. 1511-1549.

211. Elias, A.D., Triple-negative breast cancer: a short review. American journal of clinical oncology, 2010. 33(6): p. 637-645.

212. Matsuda, N., et al., Early clinical development of epidermal growth factor receptor targeted therapy in breast cancer. Expert opinion on investigational drugs, 2017. 26(4): p. 463-479.

213. Weng, T.-H., et al., RON and MET co-overexpression are significant pathological characteristics of poor survival and therapeutic targets of tyrosine kinase inhibitors in triple-negative breast cancer. Cancer research and treatment: official journal of Korean Cancer Association, 2020. 52(3): p. 973.

214. Xiang, H., et al., Targeting autophagy-related protein kinases for potential therapeutic purpose. Acta Pharmaceutica Sinica B, 2020. 10(4): p. 569-581.

215. Ross, J.S. and L.M. Gay, Comprehensive genomic sequencing and the molecular profiles of clinically advanced breast cancer. Pathology, 2017. 49(2): p. 120-132.

216. Nowacka-Zawisza, M. and W.M. Krajewska, Triple-negative breast cancer: molecular characteristics and potential therapeutic approaches. Postepy higieny i medycyny doswiadczalnej (Online), 2013. 67: p. 1090-1097.

217. Vijay, S. and T.S. Gujral, Non-linear Deep Neural Network for Rapid and Accurate Prediction of Phenotypic Responses to Kinase Inhibitors. Iscience, 2020. 23(5): p. 101129.

218. Damaskos, C., et al., Triple-negative breast cancer: The progress of targeted therapies and future tendencies. Anticancer research, 2019. 39(10): p. 5285-5296.

219. Tolba, M., et al., Novel combinatorial strategies for boosting the efficacy of immune checkpoint inhibitors in advanced breast cancers. Clinical and Translational Oncology, 2021: p. 1-16.

220. Ge, X., et al., EGFR tyrosine kinase inhibitor HS-10296 induces autophagy and apoptosis in triplenegative breast cancer MDA-MB-231 cells. Nan Fang yi ke da xue xue bao= Journal of Southern Medical University, 2020. 40(7): p. 981-987.

221. Kawai, M., et al., Midostaurin preferentially attenuates proliferation of triple-negative breast cancer cell lines through inhibition of Aurora kinase family. Journal of biomedical science, 2015. 22(1): p. 1-10.

222. You, K.S., et al., Dual Inhibition of AKT and MEK Pathways Potentiates the Anti-Cancer Effect of Gefitinib in Triple-Negative Breast Cancer Cells. Cancers, 2021. 13(6): p. 1205.

223. Miller, M.A., R.J. Sullivan, and D.A. Lauffenburger, Molecular pathways: receptor ectodomain shedding in treatment, resistance, and monitoring of cancer. Clinical Cancer Research, 2017. 23(3): p. 623-629.

224. Duncan, J.S., et al., Dynamic reprogramming of the kinome in response to targeted MEK inhibition in triple-negative breast cancer. Cell, 2012. 149(2): p. 307-321.

225. Haga, Y., et al., Inhibition of Akt/mTOR pathway overcomes intrinsic resistance to dasatinib in triple-negative breast cancer. Biochemical and Biophysical Research Communications, 2020. 533(4): p. 672-678.

226. Lux, M.P., et al., Update Breast Cancer 2020 Part 5–Moving Therapies From Advanced to Early Breast Cancer Patients. Geburtshilfe und Frauenheilkunde, 2021. 81(04): p. 469-480.

227. Malhotra, M.K. and L.A. Emens. The evolving management of metastatic triple negative breast cancer. in Seminars in oncology. 2020. Elsevier.

228. Verma, N., et al., Synthetic lethal combination targeting BET uncovered intrinsic susceptibility of TNBC to ferroptosis. Sci Adv, 2020. 6(34).

229. Liu, Z., et al., Systematic Analysis of the Aberrances and Functional Implications of Ferroptosis in Cancer. Iscience, 2020. 23(7): p. 101302.

230. Kubli, S.P., et al., AhR controls redox homeostasis and shapes the tumor microenvironment in BRCA1-associated breast cancer. Proc Natl Acad Sci U S A, 2019. 116(9): p. 3604-3613.

231. Azimi, I., et al., Hypoxia-induced reactive oxygen species mediate N-cadherin and SERPINE1 expression, EGFR signalling and motility in MDA-MB-468 breast cancer cells. Sci Rep, 2017. 7(1): p. 15140.

232. Roux, C., et al., Reactive oxygen species modulate macrophage immunosuppressive phenotype through the up-regulation of PD-L1. Proc Natl Acad Sci U S A, 2019. 116(10): p. 4326-4335.

233. Kwon, Y., Possible Beneficial Effects of N-Acetylcysteine for Treatment of Triple-Negative Breast Cancer. Antioxidants (Basel), 2021. 10(2).

234. Huang, K., et al., Multi-Omics Perspective Reveals the Different Patterns of Tumor Immune Microenvironment Based on Programmed Death Ligand 1 (PD-L1) Expression and Predictor of Responses to Immune Checkpoint Blockade across Pan-Cancer. Int J Mol Sci, 2021. 22(10).

235. Wang, H., et al., A pan-cancer perspective analysis reveals the opposite prognostic significance of CD133 in lower grade glioma and papillary renal cell carcinoma. Sci Prog, 2021. 104(2): p. 368504211010938.

236. Chiu, Y.C., et al., Deep learning of pharmacogenomics resources: moving towards precision oncology. Brief Bioinform, 2020. 21(6): p. 2066-2083.

237. Shen, L., et al., Role of PRDM1 in Tumor Immunity and Drug Response: A Pan-Cancer Analysis. Front Pharmacol, 2020. 11: p. 593195.

238. Dzneladze, I., et al., SubID, a non-median dichotomization tool for heterogeneous populations, reveals the pan-cancer significance of INPP4B and its regulation by EVI1 in AML. PLoS One, 2018. 13(2): p. e0191510.

239. Liu, Z. and S. Zhang, Tumor characterization and stratification by integrated molecular profiles reveals essential pan-cancer features. BMC Genomics, 2015. 16(1): p. 503.

240. Liu, L., et al., Combination of TMB and CNA Stratifies Prognostic and Predictive Responses to Immunotherapy Across Metastatic Cancer. Clin Cancer Res, 2019. 25(24): p. 7413-7423.

241. Ho, K.H., et al., Glycolysis-associated lncRNAs identify a subgroup of cancer patients with poor prognoses and a high-infiltration immune microenvironment. BMC Med, 2021. 19(1): p. 59.

242. Sun, W., et al., Association between Socioeconomic Status and One-Month Mortality after Surgery in 20 Primary Solid Tumors: a Pan-Cancer Analysis. J Cancer, 2020. 11(18): p. 5449-5455.

243. Wang, S., et al., Incidence and prognosis of liver metastasis at diagnosis: a pan-cancer population-based study. Am J Cancer Res, 2020. 10(5): p. 1477-1517.

244. Yan-Fei, H., et al., Dysregulation in nucleic acid-sensing pathway genes is associated with cancer patients' prognosis. Cancer Sci, 2020. 111(7): p. 2212-2222.

245. Vellichirammal, N.N., et al., Pan-Cancer Analysis Reveals the Diverse Landscape of Novel Sense and Antisense Fusion Transcripts. Mol Ther Nucleic Acids, 2020. 19: p. 1379-1398.

246. Dayton, J.B. and S.R. Piccolo, Classifying cancer genome aberrations by their mutually exclusive effects on transcription. BMC Med Genomics, 2017. 10(Suppl 4): p. 66.

247. Zhou, W., et al., Identification of driver copy number alterations in diverse cancer types and application in drug repositioning. Mol Oncol, 2017. 11(10): p. 1459-1474.

248. Lv, Y., et al., Landscape of cancer diagnostic biomarkers from specifically expressed genes. Brief Bioinform, 2020. 21(6): p. 2175-2184.

249. Wang, Y., et al., DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration. Briefings in Bioinformatics, 2021.

250. Wang, J., et al., Pathway-Based Drug Repurposing with DPNetinfer: A Method to Predict Drug–Pathway Associations via Network-Based Approaches. Journal of Chemical Information and Modeling, 2021. 61(5): p. 2475-2485.

251. Sengupta, S., et al., Integrative omics analyses broaden treatment targets in human cancer. Genome Med, 2018. 10(1): p. 60.

252. Auyeung, K.K., P.C. Law, and J.K. Ko, Combined therapeutic effects of vinblastine and Astragalus saponins in human colon cancer cells and tumor xenograft via inhibition of tumor growth and proangiogenic factors. Nutr Cancer, 2014. 66(4): p. 662-74.

253. Jaferian, S., M. Soleymaninejad, and H. Daraee, Verapamil (VER) Enhances the Cytotoxic Effects of Docetaxel and Vinblastine Combined Therapy Against Non-Small Cell Lung Cancer Cell Lines. Drug Res (Stuttg), 2018. 68(3): p. 146-152.

254. Boven, E., et al., The anti-tumour effects of the prodrugs Nl-leucyl-doxorubicin and vinblastine-isoleucinate in human ovarian cancer xenografts. British journal of cancer, 1992. 66(6): p. 1044-1047.

255. Kim, J.S., et al., The Impact of Cetuximab Plus AKT- or mTOR- Inhibitor in a Patient-Derived Colon Cancer Cell Model with Wild-Type RAS and PIK3CA Mutation. J Cancer, 2017. 8(14): p. 2713-2719.

256. He, K., et al., BRAFV600E-dependent Mcl-1 stabilization leads to everolimus resistance in colon cancer cells. Oncotarget, 2016. 7(30): p. 47699-47710.

257. Xiang, X., et al., Everolimus inhibits the proliferation and migration of epidermal growth factor receptor-resistant lung cancer cells A549 via regulating the microRNA-4328/phosphatase and tensin homolog signaling pathway. Oncol Lett, 2019. 18(5): p. 5269-5276.

258. Guo, H., et al., Everolimus exhibits anti-tumorigenic activity in obesity-induced ovarian cancer. Oncotarget, 2016. 7(15): p. 20338-56.

259. Sun, J., et al., The multi-targeted kinase inhibitor sunitinib induces apoptosis in colon cancer cells via PUMA. PLoS One, 2012. 7(8): p. e43158.

260. Mahalingam, D., et al., Heightened JNK Activation and Reduced XIAP Levels Promote TRAIL and Sunitinib-Mediated Apoptosis in Colon Cancer Models. Cancers (Basel), 2019. 11(7).

261. Gridelli, C., et al., Sorafenib and sunitinib in the treatment of advanced non-small cell lung cancer. Oncologist, 2007. 12(2): p. 191-200.

262. Socinski, M.A., The current status and evolving role of sunitinib in non-small cell lung cancer. J Thorac Oncol, 2008. 3(6 Suppl 2): p. S119-23.

263. Mologni, L., et al., Valproic acid enhances bosutinib cytotoxicity in colon cancer cells. Int J Cancer, 2009. 124(8): p. 1990-6.

264. Fowler, A.J., et al., Multikinase Abl/DDR/Src Inhibition Produces Optimal Effects for Tyrosine Kinase Inhibition in Neurodegeneration. Drugs R D, 2019. 19(2): p. 149-166.

265. Munshi, A., et al., Vorinostat, a histone deacetylase inhibitor, enhances the response of human tumor cells to ionizing radiation through prolongation of γ-H2AX foci. Molecular cancer therapeutics, 2006. 5(8): p. 1967-1974.

266. Wishart, D.S., et al., DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research, 2018. 46(D1): p. D1074-D1082.

267. Sung, G.J., et al., Inhibition of TFEB oligomerization by co-treatment of melatonin with vorinostat promotes the therapeutic sensitivity in glioblastoma and glioma stem cells. J Pineal Res, 2019. 66(3): p. e12556.

268. Lin, C.Y., et al., Vorinostat combined with brigatinib overcomes acquired resistance in EGFR-C797S-mutated lung cancer. Cancer Lett, 2021. 508: p. 76-91.

269. Gray, J.E., et al., Phase I/Ib Study of Pembrolizumab Plus Vorinostat in Advanced/Metastatic Non-Small Cell Lung Cancer. Clin Cancer Res, 2019. 25(22): p. 6623-6632.

270. Park, S.E., et al., Vorinostat enhances gefitinib-induced cell death through reactive oxygen species-dependent cleavage of HSP90 and its clients in non-small cell lung cancer with the EGFR mutation. Oncol Rep, 2019. 41(1): p. 525-533.

271. Rodriguez, G.C., et al., Phase II Trial of Chemopreventive Effects of Levonorgestrel on Ovarian and Fallopian Tube Epithelium in Women at High

Risk for Ovarian Cancer: An NRG Oncology Group/GOG Study. Cancer Prev Res (Phila), 2019. 12(6): p. 401-412.

272. Sano, K., et al., [A preliminary study to determine the activity of topoisomerase II in human kidney cancer cells with DNA unknotting method]. Nihon Hinyokika Gakkai Zasshi, 1992. 83(9): p. 1511-6.

273. Ishii, H., et al., Atezolizumab plus carboplatin and etoposide in small cell lung cancer patients previously treated with platinum-based chemotherapy. Invest New Drugs, 2021. 39(1): p. 269-271.

274. Goldman, J.W., et al., Patient-reported outcomes with first-line durvalumab plus platinum-etoposide versus platinum-etoposide in extensive-stage small-cell lung cancer (CASPIAN): a randomized, controlled, open-label, phase III study. Lung Cancer, 2020. 149: p. 46-52.

275. Paz-Ares, L., et al., Durvalumab plus platinum-etoposide versus platinum-etoposide in first-line treatment of extensive-stage small-cell lung cancer (CASPIAN): a randomised, controlled, open-label, phase 3 trial. Lancet, 2019. 394(10212): p. 1929-1939.

276. Bouzo, B.L., et al., Sphingomyelin nanosystems loaded with uroguanylin and etoposide for treating metastatic colorectal cancer. Sci Rep, 2021. 11(1): p. 17213.

277. Luo, L.X., et al., Identification of mitoxantrone as a new inhibitor of ROS1 fusion protein in non-small cell lung cancer cells. Medchemcomm, 2017. 8(3): p. 621-624.

278. Ge, C., et al., Suppression of oxidative phosphorylation and IDH2 sensitizes colorectal cancer to a naphthalimide derivative and mitoxantrone. Cancer Lett, 2021. 519: p. 30-45.

279. Guo, J., et al., Low Expression of Smurf1 Enhances the Chemosensitivity of Human Colorectal Cancer to Gemcitabine and Cisplatin in Patient-Derived Xenograft Models. Transl Oncol, 2020. 13(9): p. 100804.

280. Park, J.Y., et al., Gemcitabine-Incorporated G-Quadruplex Aptamer for Targeted Drug Delivery into Pancreas Cancer. Mol Ther Nucleic Acids, 2018. 12: p. 543-553.

281. Burris, H.A., 3rd, et al., Improvements in survival and clinical benefit with gemcitabine as first-line therapy for patients with advanced pancreas cancer: a randomized trial. J Clin Oncol, 1997. 15(6): p. 2403-13.

282. Samanta, A., G. Ravindran, and A. Sarkar, Quinacrine causes apoptosis in human cancer cell lines through caspase-mediated pathway and regulation of small-GTPase. J Biosci, 2020. 45.

VITA

Zainab Al-Taie was born in Baghdad, Iraq. She earned her bachelor's degree in Computer Science from the College of Science, the University of Baghdad in Iraq. She worked as a programmer and lecturer in the same institute for five years. She was awarded a scholarship from the Iraqi prime minister's office to pursue a master's degree in computer science.

She earned her master's degree in Computer Science from the University of Missouri-Columbia, USA. During the master's study, her research focused on developing algorithms and tools to reduce the complexity of biological networks and enable precision medicine. Then, she started her Ph.D. study and worked as a graduate research assistant in the Institute of Data Science and Informatics, University of Missouri-Columbia, USA. Her research work during the Ph.D. continued to serve the purpose of implementing precision medicine in the healthcare system by developing a novel Explainable Artificial Intelligence (XAI) method for patient stratification and drug repositioning to improve patient outcomes.

She served as the president of the Institute for Data Science and Informatics Graduate Student Organization from 2018 to 2019. Also, she served as the representative of the Institute at the Graduate Professional Council from 2017 to 2021. Nationally, she served as a reviewer in conferences. Internationally, she served as a reviewer and a member of the organization committee for the Network Biology Community of Special Interest (NetBio COSI) Track at Intelligent Systems for Molecular Biology (ISMB) conference at Basel, Switzerland in 2019 and Montreal, Canada, in 2020. She accepted an offer to join Icahn School of Medicine at Mount Sinai in New York, the USA as a postdoctoral fellow.