PANACEA: Towards Influence-driven Profiling of Drug Target Combinations in Cancer Signaling Networks

Baihui Xu
Nanyang Technological University
Singapore
nerissa.xu@ntu.edu.sg

Sourav S Bhowmick
Nanyang Technological University
Singapore
assourav@ntu.edu.sg

Jiancheng Hu
National Cancer Centre Singapore
Singapore
hu.jiancheng@nccs.com.sg

Abstract

In this paper, we introduce a novel framework called PANACEA, designed to *profile* known cancer target combinations in cancer type-specific signaling networks. Given a large signaling network for a cancer type, known targets from approved anticancer drugs, a set of cancer mutated genes, and a *combination size parameter* k, PANACEA automatically generates a *delta histogram* that depicts the distribution of k-sized target combinations based on their *topological influence* on cancer mutated genes and other nodes. To this end, we formally define the novel problem of *influence-driven target combination profiling* (i-TCP) and propose an algorithm that employs two innovative personalized PageRank-based measures, *PEN distance* and *PEN-diff*, to quantify this influence and generate the delta histogram. Our experimental studies on signaling networks related to four cancer types demonstrate that our proposed measures outperform several popular network properties in profiling known target combinations. Notably, we demonstrate that PANACEA can significantly reduce the candidate k-node combination exploration space, addressing a longstanding challenge for tasks such as *in silico* target combination prediction in large signaling networks.

CCS Concepts

• **Applied computing** → *Systems biology*.

Keywords

Data profiling, cancer signaling networks, personalized PageRank, target combinations

ACM Reference Format:
Baihui Xu, Sourav S Bhowmick, and Jiancheng Hu. 2024. PANACEA: Towards Influence-driven Profiling of Drug Target Combinations in Cancer Signaling Networks. In *15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '24), November 22–25, 2024, Shenzhen, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3698587.3701540

1 Introduction

Enhancing the quality of target selection is generally regarded as the most crucial factor for boosting productivity in the pharmaceutical industry [6]. This has led to a focus on identifying effective target combinations for specific diseases (a.k.a *multi-target selection* [9]) which presents significant challenges [6, 9, 11]. Consequently, there is growing research on *in silico* techniques to predict these combinations [2, 4, 5, 10, 14, 15, 17]. However, current studies have two main limitations. First, they neglect the *profiles* of *approved* target combinations for specific diseases to guide discovery. Second, they are mainly suited for small networks, making it difficult to comprehensively explore larger candidate combination spaces in large networks. This paper presents a novel framework designed to *profile* known target combinations in cancer signaling networks, which has the potential to address these limitations.

The *Cancer Drugs Database*, maintained by the *Anticancer Fund* [12], offers a curated list of anticancer drugs approved by one or more regulatory agencies (such as the FDA or NCI) for various cancer types, along with their associated target combinations. In this paper, we aim to *profile*[1] these target combinations in the signaling network of a specific cancer type. These profiles can potentially serve as a guide for determining which candidate k-node combinations to further analyze for target combination prediction (*i.e.,* k molecules that can be targeted simultaneously). For example, one might choose to investigate network regions where most or very few known target combinations are located.

Profiling known target combinations in signaling networks presents a significant challenge. Since various profiles can be generated, not all are effective for characterizing these combinations to address downstream problems such as target combination prediction. For example, degree or PageRank distributions may fail to capture the off-target effects of these combinations, which are a primary reason for drug failures [16]. Therefore, it is crucial to select appropriate features for profiling that can support the downstream analyses. Additionally, to manage large and noisy cancer-specific signaling networks, the chosen features must be purely topological. That is, they should not rely on complete mathematical models of the underlying cancer network, nor on the pharmacological or chemical properties of all involved molecules, or the genomic and proteomic data of patients. Instead, the profiling strategy must adopt a realistic perspective, acknowledging that such data is often unavailable in many areas of cancer signaling networks.

In this paper, we introduce a novel *influence-driven target combination profiling* (i-TCP) problem to tackle these challenges. Given a large signaling network G_C for a cancer type C (*e.g.,* breast, colorectal), known targets tackled by anticancer drugs of C, a set of cancer mutated genes (*e.g.,* oncogenes) in G_C, and a *combination size parameter* $k > 1$, the goal of i-TCP is to generate a *delta histogram* that depicts the distribution of k-size known target combinations in

BCB '24, November 22–25, 2024, Shenzhen, China

ACM ISBN 979-8-4007-1302-6/24/11
https://doi.org/10.1145/3698587.3701540

[1] *Data profiling* is the set of activities and processes to determine the metadata of a given dataset [1].

G_C based on their *topological influence* (*i.e.,* strength of connection) on cancer mutated genes and rest of the nodes in G_C.

We present a novel framework called PANACEA (**P**ersonalized p**A**gerank-based profili**N**g of c**An**C**E**r t**A**rgets) to address the *i*-TCP problem. This framework utilizes a new personalized PageRank-based (PPR) [20] measure called *PEN distance* to assess the topological influence of one node (*e.g.,* a drug target) on another (*e.g.,* an oncogene) in a cancer signaling network. Intuitively, a smaller PEN distance suggests a greater influence due to a higher number of paths between the node pair. Building on PEN distance, we also propose a measure called *PEN-diff*, which captures the difference in average influence of a k-node combination on a set of cancer mutated genes compared to the remaining nodes in the network. A positive PEN-diff value indicates that the average influence on the cancer mutated genes is greater than on other nodes, effectively capturing the off-target effects of a drug from a topological perspective (*i.e.,* impact on rest of the nodes in G_C). Using the PEN-diff values of k-node combinations and known target combinations in C, we build a *delta histogram* to depict the distribution of known k-target combinations in the PEN-diff space of G_C. We experimentally demonstrate that PANACEA outperforms two baseline strategies in profiling target combinations across four different cancer types.

2 Background

In this section, we provide background information to facilitate exposition of PANACEA.

Human Signaling Network. Large signaling networks often are modeled simply as large graphs. For instance, Cui *et al.* [8] modeled the human signaling network as a graph where nodes represent proteins. Directed links are used to represent activation or inhibition whereas undirected links represent physical interactions of proteins that are not characterized as activating or inhibitory. There are two types of directed links, incoming and outgoing. The incoming link represents a signaling from another node whereas an outgoing link represents a signal to another node. These two types of directed links are collectively referred to as *signal links*. In contrast, the physical links are referred to as *neutral links*. In this paper, we represent the human signaling network using this model.

In our work, we utilize the signaling network used in the study reported in [21]. It contains 6,305 nodes and 62,937 links. There are 33,398 activation links (*i.e.,* positive links), 7,960 inhibitory links (*i.e.,* negative links), and 21,579 physical links (neutral links). Since we focus on signal links, we reduce this network by removing all neutral links. This resulted in the *reduced* network with 6,009 nodes and 41,358 edges.

Cancer Mutated Genes. We gather cancer mutated genes from the *COSMIC Cancer Gene Census* database [13]. These genes can be categorized into three types, positive regulators (oncogenes), negative regulators (tumor suppressors), and fusion genes. Since majority of the mutated genes in cancer are oncogenes [8], we focus only on them in this work. In our dataset, there are 318 oncogenes, 252 (79.25%) of which can be located in the reduced network.

Cancer Drug Targets. We obtain data on drug targets from the *Cancer Drugs Database* maintained by *Anticancer Fund* [12]. Among the 1,025 drug targets, 119 (11.61%) are oncogenes and 725 (70.73%) are in the human signaling network.

Personalized PageRank (PPR) [20]. Let $G = (V, E)$ be a directed graph where V is the set of nodes (vertices) and E is the set of edges (links). Given a source node $s \in V$ and a *jump factor* α, a walker starts from s to traverse G, and at each step, the walker either (a) terminates at the current node with a probability α, or (b) jumps to a randomly selected out-neighbor of the current node. For any node $t \in V$, the *personalized PageRank* (PPR) $\pi(s, t)$ is the probability that a random walk (RW) from s terminates at t. Intuitively, a large $\pi(s, t)$ indicates that many paths exist from s to t. That is, s is well connected to t. The α value is usually set to 0.15 or 0.2 [20, 22].

3 The *i*-TCP Problem

Intuitively, the goal of the *influence-driven target combination profiling* (*i*-TCP) problem is to generate a *delta histogram* that visually represents the distribution of known k-target combinations w.r.t. their *topological influence* in a cancer signaling network. Given a large signaling network $G_C = (V_C, E_C)$ for a cancer type C (*e.g.,* breast, colorectal), known targets $X_t \subseteq V_C$ of C, a set of cancer mutated genes $X_g \subseteq V_C$ in G_C, and a k-combination node set h_k where $k > 1$ and $h_k \subset V_C$, the *topological influence* of h_k is defined by a function $f(h_k, X_g, G_C)$ that quantifies the *aggregate* influence of h_k on X_g and $(V_C - X_g)$.

A *delta histogram* $\mathbb{H}_C$ of G_C is an equi-width histogram whose X-axis represents N_{bucket} buckets and the Y-axis represents the *percentages* of known k-target combinations in G_C in each bucket. A k-node combination h_k is assigned to a bucket $b_i = [r_{min}, r_{max}]$ if $r_{min} < f(h_k, X_g, G_C) \leq r_{max}$. The *percentage* of known k-target combinations in b_i is the number of known target combinations in the top-m percentage of k-node combinations for $m > 0$.

*Definition 3.1. Given a signaling network $G_C = (V_C, E_C)$ for a cancer type C, a set of known targets $X_t \subseteq V_C$, a set of cancer-mutated genes $X_g \subseteq V_C$, a combination size parameter $k > 1$, and the number of buckets $N_{bucket} > 1$, the goal of **influence-driven target combination profiling** problem is to compute the following:*

$$\mathbb{H}_{k,N_{bucket}} = \mathcal{F}(\mathcal{G}(G_C, k, X_g), X_t, N_{bucket}) \quad (1)$$

where $\mathcal{G}(G_C, k, X_g)$ is a function that computes $f(h_k, X_g, G_C)$ of the k-combination nodes in G_C and $\mathcal{F}(\cdot)$ is a function that computes an N_{bucket}-delta histogram of the known k-target combinations in X_t.

In PANACEA the topological influence function $f(\cdot)$ is realized by a novel PPR-based measure called *PEN-Diff* that captures the interplay between the influence of a k-target combinations on X_g and rest of the network (*i.e.,* off-target effect).

4 The PANACEA Framework

In this section, we present the PANACEA framework to address the *i*-TCP problem in cancer signaling networks. We begin by introducing a PPR-based node distance measure called *PEN distance* that serves as the foundation for this task.

4.1 PEN Distance

While several studies in drug and target combination discovery utilize various distance-based measures [3, 10], it is important to recognize that the shortest path is not the only way a target can impact other nodes. Targets may connect to other nodes via multiple alternative paths. Therefore, assessing topological influence should

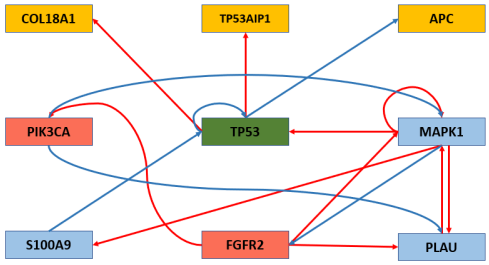

Figure 1: [Best viewed in color] Target-aware cancer-specific signaling network for breast cancer. The red edges are positive links and the blue edges are negative links. Nodes in yellow are drug targets. Nodes in green are cancer genes. Red nodes are both a cancer gene and a drug target.

leverage network propagation-based methods. Cowen *et al.* [7] have reviewed several of these methods, which include methods based on random walk with restart (*e.g.,* PPR). In our work, we focus on capturing connectivity-based relationships through a PPR-based distance measure known as *PEN distance* (**P**ersonalized pag**E**rank-based **N**ode distance).

Intuitively, the PPR value $\pi(s,t)$ indicates the node t's importance to the node s. Recall that if $\pi(s,t)$ is high then t can be reached from s via many paths (*i.e.,* t is important w.r.t. s as it can be reached by many paths from s). Our goal is to capture this importance between a pair of nodes (*e.g.,* a drug target FGFR2 and a tumor suppressor TP53 in Figure 1) in a cancer signaling network in the form of a "distance" measure. That is, if a node pair (s,t) has a high PPR value then the "distance" should be small. Unfortunately, the PPR values of adjacent node pairs in a graph could vary significantly [22]. Consequently, directly using the PPR values as a "distance" measure may inject a large variance in their values between the adjacent nodes (*e.g.,* adjacent drug targets) and other nodes (*e.g.,* oncogenes). To alleviate this challenge, we propose *PEN distance* as follows.

Definition 4.1. **[PEN Distance]** *Given a signaling network $G = (V,E)$, the PEN distance between nodes $s \in V$ and $t \in V$ is defined as follows:*

$$P[s,t] = \begin{cases} 0 & \text{if } s = t\,, \\ 1 - \log(\pi_d(s,t) + \epsilon) & \text{if } s \neq t\,. \end{cases} \tag{2}$$

where $s \neq t$, $\pi_d(s,t) = \pi(s,t) \times d(s)$, $d(s)$ is the out-degree of node s, and $\epsilon = 1e-5$ where e is the Euler constant.

In the above definition, $\pi_d(s,t)$ is the *degree-normalized PPR* (DPPR) from s to t. According to [19], multiplying a node's PPR by its out-degree yields a more precise measure of the strength of connections between nodes. Additionally, employing DPPR helps mitigate the variability of PPR among neighboring nodes. The ϵ parameter is added to avoid undefined $log(\cdot)$ value when $\pi(s,t) = 0$. Intuitively, if $\pi(s,t)$ is large, then $P[s,t]$ tends to be small, *i.e.,* well-connected nodes have a closer distance from each other.

4.2 PEN-Diff

Next, we introduce the notion of *PEN-diff*, which is based on PEN distance. Let $V_g \subset V_C$ be a set of nodes in a signaling network G_C.

Table 1: Target-aware cancer-specific signaling networks.

Type	No. of Nodes	No. of Edges	No. of Known Targets
Prostate Cancer	2,214	27,899	87
Breast Cancer	2,560	29,986	295
Bladder Cancer	2,291	28,602	93
Colorectal Cancer	2,467	29,717	109

Consider a node $s \in V_C$. The *average PEN distance* from s to V_g, denoted as $\overline{P}[s,V_g]$, is given as follows: $\overline{P}[s,V_g] = \frac{\sum_{i=1}^{|V_g|} P[s,t_i]}{|V_g|}$.

Then, the *single-source PEN-diff* of a node s in $G_C = (V_C, E_C)$, denoted as $P_\Delta[s,V_g]$, is the difference between the average PEN distance to the the nodes in $V_C - V_g$ and the average PEN distance to the nodes in V_g. Formally, it is defined as follows: $P_\Delta[s,V_g] = \overline{P}[s,V_C - V_g] - \overline{P}[s,V_g]$.

Given two sets of nodes V_s and V_g, the *PEN-diff* of V_s is the average single-source PEN-diff values of the nodes in V_s w.r.t. V_g and rest of the nodes in the network. Formally,

Definition 4.2. **[PEN-diff]** *Given $V_s \subset V_C$ and $V_g \subset V_C$ in a signaling network $G_C = (V_C, E_C)$, the PEN-diff of V_s is defined as follows.*

$$\overline{P}_\Delta[V_s,V_g] = \frac{\sum_{i=1}^{|V_s|} P_\Delta[v_i,V_g]}{|V_s|} \tag{3}$$

Remark. Observe that if $\overline{P}_\Delta[V_s,V_g] > 0$ then the average PEN distance of nodes in V_s and $(V_C - V_g)$ is larger than the average PEN distance with V_g. That is, the nodes in V_s are relatively less connected to the nodes in $(V_C - V_g)$ compared to the nodes in V_g. In the next subsection, we shall represent a set of oncogenes using V_g and target combinations using V_s. Consequently, a positive PEN-diff value for a target combination set indicates that these targets exert relatively less influence on the rest of the network compared to the oncogenes, which is desirable due to off-target effects. Note that PANACEA is flexible to represent other types of nodes (*e.g.,* biomarkers, disease-related nodes) as V_g.

4.3 Profiling Drug Target Combinations

We now present the algorithm to profile the known target combinations in a cancer signaling network by exploiting PEN distance. The formal description of the algorithm is given in [18].

Phase 1: Target-aware cancer-specific signaling network construction. Given that cancer is a complex disease involving different genes and proteins for various types (*e.g.,* breast, colorectal) and that drug targets differ by cancer type, we first extract a subnetwork from the reduced human signaling network corresponding to a specific cancer type C and the known targets for C. Given the reduced signaling network $G = (V,E)$, a cancer type C, a set of known targets X_t for C from the *Anticancer Fund* database [12], a set of oncogenes X_g from *COSMIC Cancer Gene Census* database [13], and a configurable *path length threshold* d ($d = 5$ by default), the algorithm searches for paths with length less than d between each pair of a known target $u \in X_t$ and an oncogene $v \in X_g$ in G. The nodes in each such path $p(u,v)$ and their edges are added to a path set P_C and then used to extract the output subnetwork G_C from G. That is, the subnetwork encompasses all nodes associated with the oncogenes and targets for C, along with the intermediate nodes that link them. Figure 1 shows a fragment of the network constructed

Table 2: Example of PEN distance matrix of selected nodes in Figure 1.

source node	target node TP53	PIK3CA	S100A9	TP53AIP1	FGFR2	APC	COL18A1	MAPK1	PLAU
TP53AIP1	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129
FGFR2	2.4265	2.6688	3.2374	4.0358	1.3471	4.0358	4.0358	1.4049	1.9927
APC	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129
PIK3CA	2.763	1.8667	3.5739	4.3722	3.5739	4.3722	4.3722	1.7414	2.3292
COL18A1	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129	12.5129

Table 3: Example of PEN-diff computation.

2-node pair	$\overline{P}[V_s, V_C - V_g]$	$\overline{P}[V_s, V_g]$	PEN-Diff
(FGFR2, PIK3CA)	3.1732	2.4313	0.7419
(FGFR2, TP53AIP1)	7.6914	7.5303	0.1611

by the algorithm for breast cancer. Table 1 reports the features of target-aware, cancer-specific signaling networks for four different cancer types.

Phase 2: PEN distance computation. Given the constructed target-aware cancer-specific signaling network G_C, this phase computes the PEN distance between all pairs of nodes in the network by exploiting Definition 4.1. The key step here is the computation of PPR values of pairs of nodes. Since this phase incurs a one-time cost for a specific signaling network, we calculate the exact PPR values. It produces the *PEN distance matrix* **P**, which contains the PEN distance values for all pairs of nodes in G_C.

Consider Figure 1. The PEN distance matrix involving some of the node pairs is shown Table 2. Observe that there are multiple paths from various nodes to PIK3CA, resulting in low PEN distance between these pairs. In contrast, TP53AIP1 and the other nodes exhibit high distance, as there are very few (if any) paths linking them to TP53AIP1.

Phase 3: PEN-diff computation phase In this phase, we utilize the PEN distance matrix **P** to compute the PEN Diff (Definition 4.2) for k-node combinations in G_c w.r.t. the oncogenes and non-oncogenes. Consider the set of oncogenes X_g in G_c and a *combination size parameter* k ($k > 1$). For each k-nodes pair of V_c, denoted as $H_{k,i}$, we compute the *average* PEN-distance of the nodes in $H_{k,i}$ with the nodes in X_g. Similarly, we calculate the average PEN-distance with the nodes in $V_C - X_g$. Finally, the difference between these two values is used to compute the PEN-diff of $H_{k,i}$, which is then stored in the PEN-diff hash table D.

The PEN-diff values of two 2-node combinations in Figure 1 are shown in Table 3. Notably, the PEN-diff values are positive. The nodes FGFR2 and PIK3CA are both targets of the drug *Fulvestrant*, a hormone treatment for advanced breast cancer. This pair has an average PEN distance of 2.4313 with oncogenes and 3.1732 with other nodes, indicating that they are, on average, significantly more connected to oncogenes than to the rest of the network. In contrast, the nodes FGFR2 and TP53AIP1 are not targets of the *same* drug (*i.e.,* not a target combination), resulting in a considerably lower PEN-diff value of 0.1611.

Phase 4: Target combination profiling. The PEN-diff hash table D generated in the preceding phase is exploited in this phase to profile known target combinations in G_C. Given D, a set of known targets X_t for the cancer C, and the number of buckets N_{bucket}, it first groups the k-node combinations in D according to their PEN-diff values. Specifically, this step generates equi-width N_{bucket} (by default $N_{bucket} = 5$) buckets based on the PEN-diff values and assigns each combination in D to the appropriate bucket $b_i = [r_{min}, r_{max}]$. That is, a node combination c is assigned to a bucket b_i if $D[c] > r_{min}$ and $D[c] \leq r_{max}$. Node combinations within each bucket are sorted by their PEN-diff values in descending order (*i.e.,* a lower rank indicates greater similarity in average connectivity or influence between V_g and rest of the nodes). Then the percentage of known k-target combinations in a specific bucket is computed by counting the number of known target combinations in the top-m percentage of node combinations by varying $m \in 1, 10, 20, 50$. This is used to construct the *delta-histogram.* The X-axis of the delta histogram represents the buckets, while the Y-axis indicates the percentages of known k-target combinations in each bucket for different values of m. The $[r_{min}, r_{max}]$ values of the bucket with the highest *coverage* are returned as the target profile thresholds δ_{min} and δ_{max} (ties are broken arbitrarily). We define the *coverage* of a bucket as the percentage of known target combinations in the top-50% of its node combinations. These values for a specific cancer-type network and the delta histogram can be used to prioritize k-target combinations.

Figure 2 (left) depicts the delta histogram of the breast cancer-specific signaling network in Table 1. Observe that majority of the known targets are found in the first bucket. Specifically, the bucket $[-0.0045, 0.4301]$ has the highest coverage of known target combinations for all values of m. Hence, these values are returned as δ_{min} and δ_{max}, respectively, along with the delta histogram. Additionally, most known target combinations fall within the top 1% of the first bucket, and for the remaining m values, there are no known target combinations beyond those in the top 1%. As a result, the bars in the first bucket are of equal length.

5 Performance Study

PANACEA is implemented using Python. We shall now present the key performance results of PANACEA. Additional experiments are reported in [18]. All experiments are performed on a 64-bit Windows machine with 12th Gen Intel(R) Core(TM) i7-1250U CPU(1.10 GHz) and 32.0 GB of main memory.

5.1 Experimental Setup

Datasets. We use the four types of cancer signaling networks in Table 1. We use the sets of oncogenes and known targets as detailed in Section 2. We set $\alpha = 0.2$, $k = 2$, and $N_{bucket} = 5$ for our experiments. Specifically, we set $k = 2$ because most literature on *in silico* target combination discovery primarily focuses on identifying 2 or 3 target combinations for combination therapy [5, 10, 15]. We set $m \in \{10, 20, 40, 50\}$.

Baselines. We compare PANACEA with the following baselines. (a) **PPR-diff**: Recall that a key motivation for introducing PEN distance is that we cannot effectively use PPR to profile known target combinations. Therefore, in this baseline, we use PPR values of the node pairs in G_C to calculate the average difference instead of PEN-diff. (b) **Distance-diff**: We leverage network distance between a pair of nodes s and t in G_C (*i.e.,* the length of the shortest path from

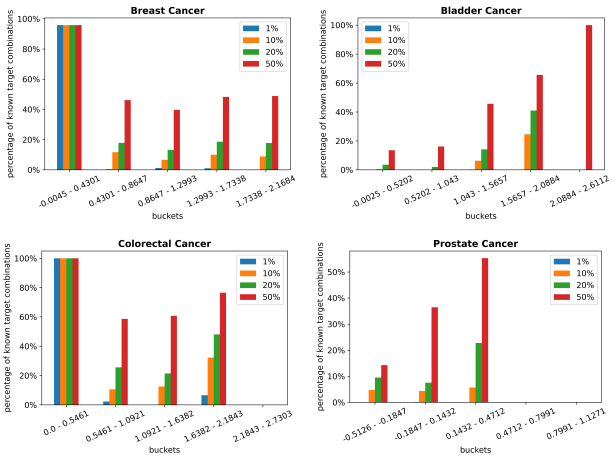

Figure 2: [Best viewed in color] Delta histograms of breast, bladder, colorectal, and prostate cancers.

Table 4: Characteristics of maximum-coverage buckets.

Type	Candidate size	Range constraint	$\|b\|$	Coverage
Breast Cancer	3, 275, 520	$[-0.0045, 0.4301]$	1, 770	95.81%
Bladder Cancer	2, 623, 195	$[2.0884, 2.6112]$	1, 985	100%
Colorectal Cancer	3, 041, 811	$[0.0, 0.5461]$	91	100%
Prostate Cancer	2, 449, 791	$[0.1432, 0.4712]$	741, 637	55.28%

s to t) in lieu of PEN distance to compute the average difference. Note that distance has been exploited by several recent techniques for target/drug combination discovery [3, 10]. Additionally, it serves as the basis for various network centrality measures, including betweenness and closeness centrality.

The computation of *PPR-diff* and *Distance-diff* is similar to PEN-diff. Given a signaling network G_C, for a k-node combination, we compute the average PPR (resp. distance) to the set to oncogenes V_g and rest of the nodes $V_C - V_g$ in the cancer-specific signaling network. We then find the difference to determine the *PPR-diff* (resp. *Distance-diff*).

5.2 Delta Histograms

We first report the delta histograms generated by PANACEA and their characteristics. Figure 2 plots the results. We can make the following observations. Firstly, the buckets with the highest coverage vary across different cancers. Specifically, for breast and colorectal cancers, the first bucket shows the maximum coverage, while for bladder and prostate cancers, the buckets $[2.0884, 2.6112]$ and $[0.1432 - 0.4712]$ have the highest coverage, respectively. This variation is expected given the complexity and heterogeneity of different cancer types. Note that these buckets include 55.28%-100% of the known target combinations. Specifically, the buckets with the highest coverage encompass the majority of known target combinations for breast, bladder, and colorectal cancers. Secondly, a significant portion of the known target combinations (with some exceptions for breast cancer) exhibit positive PEN-diff values. This suggests that these known target combinations are more closely connected to the oncogenes than to the other nodes, indicating they exert greater topological influence on the oncogenes compared to the rest. Thirdly, the known target combinations are not always grouped in buckets with high PEN-diff values (i.e., the rightmost buckets) for *every* cancer type. This creates an opportunity to investigate candidate combinations in buckets with elevated PEN-diff values for the discovery of novel target combinations.

Table 5: Exploration Size Ratio (ESR).

Cancer Types	M	$\|b_{worst,M}\|$	ESR
Breast Cancer	PEN-diff	226, 049	1.00
	PPR-diff	539, 293	2.39
	Distance-diff	862, 918	3.82
Bladder Cancer	PEN-diff	195, 974	1.00
	PPR-diff	881, 789	4.50
	Distance-diff	322, 523	1.65
Colorectal Cancer	PEN-diff	296, 818	1.00
	PPR-diff	542, 008	1.83
	Distance-diff	649, 281	2.19
Prostate Cancer	PEN-diff	122, 070	1.00
	PPR-diff	1, 131, 200	9.27
	Distance-diff	432, 114	3.54

5.3 Comparison of Exploration Space Size

In this set of experiments, we analyze the size of the exploration space and compare it to those generated by the baseline strategies.

Size of the maximum-coverage bucket. We consider the buckets that have maximum (*i.e.,* highest) coverage of known target combinations for the four types of cancer. Table 4 reports the number of candidate k-size combination in the cancer-specific signaling network, along with the range constraints for the buckets with maximum coverage, their size $|b|$ (*i.e.,* number of k-node combinations), and the coverage of known target combinations. We note that each cancer-specific signaling network contains over two million candidate target combinations, making exhaustive exploration of this space for potential target combinations computationally challenging. However, if one opts to explore the bucket with maximum coverage, the exploration space is dramatically reduced by 99.95%, 99.92%, 99.997% and 69.73% for breast cancer, bladder cancer, colorectal cancer, and prostate cancer, respectively. For instance, if one chooses the bucket $[-0.0045, 0.4301]$ for further analysis in the breast cancer-specific signaling network, they only need to examine 1, 770 out of a total of 3, 275, 520 candidate combinations in the network. This significantly reduces the exploration space and has the potential to enhance the efficiency of downstream tasks.

Comparison of worst-case bucket size. The experiment above highlights the benefit of exploring buckets with maximum coverage. Note that one may choose to analyze other buckets as well. Hence, a key question arises: *What is the worst-case size of a bucket in the delta histogram that one may need to explore, and how does it compare to the buckets generated by PPR-diff and Distance-diff?* In this set of experiments, we shed insights into this question.

We compare the worst-case size of a bucket in the delta histograms of PANACEA with the corresponding worst-case bucket sizes in the delta histograms generated by *PPR-diff* and *Distance-diff*, respectively. To this end, we define *exploration size ratio* (ESR) as follows: $ESR_X = \frac{|b_{worst,X}|}{|b_{worst,P}|}$ where $X \in \{PPR-diff, Distance-diff, \text{PANACEA}\}$, P denotes PANACEA, and $b_{worst,M}$ is the bucket with the largest number of k-node combinations in the delta histogram generated by M. Note that $ESR_{\text{PANACEA}} = 1$. The results are shown in Table 5. Observe that *ESR* consistently exceeds 1 across all cancer types for both *PPR-diff* and *Distance-diff*. This indicates that, in the worst case, significantly larger candidate combinations must be explored when the delta histograms are generated using

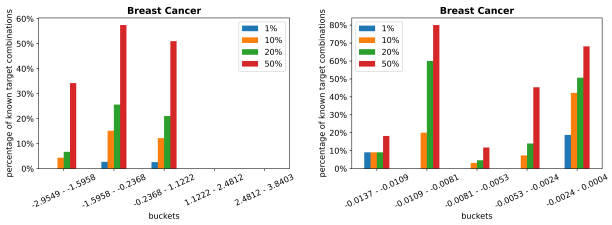

Figure 3: [Best viewed in color] Delta histograms generated by baselines for breast cancer: *Dist-diff* (left); *PPR-diff* (right).

network distance or PPR. This occurs because the *PPR-diff* and *Distance-diff* values for most candidate combinations are clustered within a narrow range [18]. Figure 3 depicts the delta histograms generated by *Dist-diff* and *PPR-diff* for breast cancer. It is evident that the coverage of the maximum-coverage buckets is lower compared to PANACEA. The results are qualitatively similar for other cancer types, indicating that these baselines are not effective for profiling known target combinations.

5.4 Usefulness of Delta Histograms

In this section, we undertake a case study to demonstrate the benefits of delta histograms generated by PANACEA. We use the results of Hu *et al.* [10] for identifying synergistic *optimal control nodes* (OCN) pairs in breast cancer. It identifies 63 synergistic OCN pairs involving 28 genes. In our breast cancer signaling network, three of these genes–PSENEN, MAML2, and GNG11–are missing. As a result, we pruned the OCN pairs involving these genes, leading to 35 synergistic OCN pairs. We found that *all* these OCN pairs are confined to just two buckets in the delta histogram for breast cancer, rather than being spread across all five buckets: $[0.8647 - 1.2993]$ and $[1.2993 - 1.7338]$ (Figure 2). Specifically, the PEN-diff values for these OCN pairs fall within $[1.1446 - 1.6398]$. *Therefore, exploring these two buckets instead of the entire breast cancer signaling network can significantly aid in the discovery of all these OCN pairs.* Furthermore, it is important to note that the PEN-diff values of all these OCN pairs are greater than one. This indicates that these OCN pairs exert relatively less influence on the rest of the network compared to the oncogenes (*i.e.,* off-target effects), which is a desirable characteristic, consistent with our PEN-diff-based model. Additionally, exploring these two buckets for target combination prediction may uncover novel target pairs with positive PEN-diff values that were not identified by Hu *et al.*

6 Related Work

There is extensive research on efficient computation of PPR with quality guarantees [20]. In this context, we investigate how two PPR-based measures, PEN distance and PEN-diff, can be utilized to profile known target combinations in cancer signaling networks. Zhang *et al.* [22] introduced a PPR-based distance measure called *PDistance*, designed to enhance graph visualization by strategically positioning nodes. While PEN distance also uses degree-normalized PPR, it has a different definition and application than PDistance.

Techniques such as random walks, random walks with restart (*e.g.,* PPR), and diffusion kernels have been used for network propagation aimed at tasks like function prediction, gene prioritization, module detection, patient stratification [7]. However, these methods typically yield scoring vectors or similarity matrices rather than delta histograms, and they do not focus on profiling known target combinations. In contrast, our research centers on profiling known target combinations in cancer signaling networks using a PPR-based approach and generates delta histograms.

7 Conclusions

This paper integrates data profiling with cancer signaling networks and drug targets by introducing the novel influence-driven target combination profiling problem and presenting PANACEA as a solution for large cancer signaling networks. It utilizes two innovative personalized PageRank-based measures, PEN distance and PEN-diff, to summarize the distribution of known targets in relation to their influence on cancer-mutated genes and other nodes in the network through delta histograms. Experimental results show that the PEN-diff-based profiling outperforms several alternative methods.

Acknowledgments

The authors are partially supported by the AcRF Tier-2 Grant MOE2019-T2-1-029.

References

[1] Z. Abedjan, L. Golab, F. Naumann, T. Papenbrock. Data Profiling. *Synthesis Lectures on Data Management*, Morgan & Claypool Publishers, 2018
[2] M. Alvarez, et al. Functional Characterization of Somatic Mutations in Cancer using Network-based Inference of Protein Activity. *Nat. Genet.* ,48, 838, 2016.
[3] F. Cheng F, I. A. Kovács, A. L. Barabási. Network-based Prediction of Drug Combinations. *Nature Communications*, 10(1), 2019.
[4] H. Chua, et al. STEROID: In silico heuristic target combination identification for disease-related signaling networks. *In ACM BCB*, 2012.
[5] H.-E. Chua, et al. Synergistic Target Combination Prediction from Curated Signaling Networks: Machine Learning meets Systems Biology and Pharmacology. *Methods*, 129, Elsevier, 2017.
[6] P. Csermely, et al. Structure and Dynamics of Molecular Networks: A Novel Paradigm of Drug Discovery: a Comprehensive Review. *Pharmacology & therapeutics*, 138(3): p. 333-408, 2013.
[7] L. Cowen et al. Network Propagation: A Universal Amplifier of Genetic Associations. *Nat Rev Genet* 18, 551–562, 2017.
[8] Q. Cui, Y. Ma et al. A Map of Human Cancer Signaling. *Molecular Systems Biology*, 3(1), John Wiley & Sons, 2007.
[9] D. Hanahan D, R. A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell* 4, 144(5):646-74, 2011.
[10] Y. Hu, C.-H. Chen, et al. Optimal Control Nodes in Disease-perturbed Networks as Targets for Combination Therapy. *Nature Communications*, 10, 2180, 2019.
[11] M. S. Mitchell. Combinations of Anticancer Drugs and Immunotherapy. *Cancer Immunology, Immunotherapy*, 52, 2003.
[12] P. Pantziarka, et al. An Open Access Database of Licensed Cancer Drugs. *Frontiers in Pharmacology*, DOI: 10.3389/fphar.2021.627574, 2021.
[13] Z. Sondka, et al. The COSMIC Cancer Gene Census: Describing Genetic Dysfunction Across all Human Cancers. *Nature Reviews Cancer*, 18(11), 2018.
[14] J. Tang, et al. Target inhibition networks: predicting selective combinations of druggable targets to block cancer survival pathways. *PLoS Comp. Bio.*, 9(9), 2013.
[15] T. D. Tran, D. T. Pham. Identification of Anticancer Drug Target Genes using an Outside Competitive Dynamics Model on Cancer Signaling Networks. *Scientific Reports*, 11, 14095, 2021.
[16] H. Vogel, et al. Drug Discovery and Evaluation: Pharmacological Assays. *Springer Science & Business Media*, 2007.
[17] X. Wang, R. Simon. Identification of Potential Synthetic Lethal Genes to p53 using a Computational Biology Approach. *BMC MedIcal Genomics*, 6, 30, 2013.
[18] B. Xu, S. S. Bhowmick, J. Hu. PANACEA: Towards Influence-driven Profiling of Drug Target Combinations in Cancer Signaling Networks. *Technical Report*, https://arxiv.org/pdf/2410.11458, Oct 2024.
[19] R. Yang, et al. Homogeneous Network Embedding for Massive Graphs via Reweighted Personalized PageRank. *PVLDB*, 13(5), 2020.
[20] M. Yang, et al. Efficient Algorithms for Personalized PageRank Computation: A Survey. *IEEE Trans. Knowl. Data Eng.* 36(9): 4582-4602, 2024.
[21] N. Zaman, L. Li, et al. Signaling Network Assessment of Mutations and Copy Number Variations Predict Breast Cancer Subtype-specific Drug Targets. *Cell Reports*, 5(1), Elsevier, 2013.
[22] S. Zhang, R. Yang, X. Xiao, X. Yan, B. Tang. PPRviz: Effective and Efficient Graph Visualization based on Personalized PageRank. *In SIGMOD*, 2023.