TENET: Topological Feature-based Target Characterization in Signaling Networks

Huey Eng Chua[§] Sourav S Bhowmick[§] Lisa Tucker-Kellogg[†]
C. Forbes Dewey, Jr.[‡]

[§]School of Computer Engineering, Nanyang Technological University, Singapore
[†]Centre for Computational Biology, Duke-NUS Graduate Medical School, Singapore
[‡]Division of Bioengineering, Massachusetts Institute of Technology, USA
CHUA0530|assourav@ntu.edu.sg, lisa.tucker-kellogg@duke-nus.edu.sg, cfdewey@mit.edu

May 22, 2015

Abstract

Target characterization for a biochemical network is a heuristic evaluation process that produces a *characterization model* that may aid in predicting the suitability of each molecule for drug targeting. These approaches are typically used in drug research to identify novel potential targets using insights from known targets. Traditional approaches that characterize targets based on their molecular characteristics and biological function require extensive experimental study of each protein and are infeasible for evaluating larger networks with poorly-understood proteins. Moreover, they fail to exploit network connectivity information which is now available from systems biology methods. Adopting a network-based approach by characterizing targets using network features provides greater insights that complement these traditional techniques. To this end, we present TENET, a network-based approach that characterizes known targets in signaling networks using topological features. TENET first computes a set of topological features and then leverages a SVM-based approach to identify *predictive topological features* that characterizes known targets. A *characterization model* is generated and it specifies which topological features are important for discriminating the targets and how these features should be combined to quantify the likelihood of a node being a target. We empirically study the performance of TENET from a wide variety of aspects, using several signaling networks from *BioModels* with real-world curated outcomes. Results demonstrate its effectiveness and superiority in comparison to state-of-the-art approaches.

1 Introduction

Complex intra- and inter-cellular signaling drives various biological processes such as growth, proliferation and apoptosis within systems. In systems biology, these molecular interactions are typically modelled as signaling networks [50] that provide a holistic view of the various interactions between different molecular players in the system. As signaling networks become an increasingly acceptable way for representing biological systems, various *network-based* computational techniques have been developed to analyze these networks with the goal of answering biological needs such as target characterization [16] and target discovery [105]. *In this paper, we focus on the target characterization problem for signaling networks.*

Target characterization identifies characteristics (*e.g.*, topological features) that distinguishes *targets* (*i.e.*, nodes) from other nodes in the network. These characteristics can be summarized as models which we refer to as *characterization models*. Traditionally, targets are characterized based on their molecular characteristics (*e.g.*, structure and binding sites of targets [62]) and biological function (*e.g.*, regulation of apoptosis [104]). These traditional approaches focus primarily on the target alone and are oblivious to the presence of other interacting molecules in the system. However, understanding how a target interacts with other molecules in a biological system may provide valuable and holistic insights for superior target characterization. For example, the degree centrality of a target may be leveraged to assess potential toxicity of targets since high degree nodes tend to be involved in essential protein-protein interactions [37] and are potentially toxic as a result. In particular, *network-based* target characterization techniques can exploit such topological features for superior characterization of targets.

Recently, there have been increasing efforts toward devising network-based target characterization techniques [41, 65, 108]. These methods focus on using topological features to characterize targets of protein-protein interaction (PPI) networks. Specifically, McDermott *et al.* performed characterization of targets in *protein co-abundance networks*[1] using several topological features such as degree centrality. Although this study suggests that multiple topological features can be combined for superior target characterization, it did not explore how these topological features should be combined towards this goal. In contrast, Hwang *et al.* concluded that *bridging centrality* is useful in identifying targets in PPI networks. However, the complexity and diversity of biological networks make target characterization using a single feature challenging since in some networks the chosen feature may perform poorly. Indeed, Chua *et al.* [16] showed that bridging centrality performs well in the `MAPK-PI3K` network, but not in the `glucose` metabolism network. Zhang *et al.* proposed the use of machine learning techniques such as support vector machines (SVM) and logistic regression for characterizing known targets in a manually curated human PPI network using 15 topological features. In contrast to [65], their goal was to identify topological characteristics of drug targets in gen-

[1]The *protein co-abundance networks* are essentially protein-protein interaction (PPI) networks constructed by identifying highly differentially regulated proteins from proteomics data using specific filters.

Symbol	Description
θ_u	Degree centrality of node u. The in, out and total degree centralities are denoted as $\theta_{in(u)}$, $\theta_{out(u)}$ and $\theta_{total(u)}$, respectively.
α_u	Eigenvector centrality of node u.
β_u	Closeness centrality of node u.
γ_u	Eccentricity centrality of node u.
δ_u	Betweenness centrality of node u.
π_u	Bridging coefficient of node u.
ζ_u	Bridging centrality of node u.
κ_u	Clustering coefficient of node u. The undirected, in, out, cycle and middleman clustering coefficients are denoted as $\kappa_{undir(u)}$, $\kappa_{in(u)}$, $\kappa_{out(u)}$, $\kappa_{cyc(u)}$ and $\kappa_{mid(u)}$, respectively.
μ_u	Proximity prestige of node u.
ω_u	Target downstream effect of node u.

Table 1: Topological features.

eral, instead of for specific diseases. However, characterizing targets in general assumes that targets of different diseases share similar target characteristics, which may not always be true. Indeed, as we shall see in Section 4, known targets in signaling networks tend to be characterized by different sets of topological features. Consequently, target characterization based on individual disease-specific networks may yield better characterization that is specific to the disease.

A common thread running through the aforementioned target characterization techniques is their focus on PPI networks. Surprisingly, similar systematic study in curated signaling networks has been lacking in the literature. Compared to signaling networks, PPI networks may contain many false-positive PPI in the sense that although these proteins can truly physically bind they may never do so inside cells due to different localization or they are not simultaneously expressed. Furthermore, PPI networks are static. That is, the edges in PPI networks are undirected; there is neither flow of information nor mass between nodes. Hence, they lack of knowledge of the underlying mechanism (*i.e.*, actual signal flow) causing the disease. Since network quality directly affects the results of network-based target characterization, the aforementioned limitations of PPI networks may adversely impact the search for superior characteristics of targets. Signaling networks, however, model the dynamic interaction of the biological systems and present an attractive alternative to PPI networks.

In our recent work [16], we took the first step to demonstrate how signaling networks can be effectively leveraged to identify topological features that are *discriminative* of targets using the Wilcoxon test. However, similar to [65], this work does not shed any insight on a *predictive model* to combine these features for identifying potential targets. In this paper, we address this limitation by presenting TENET (**T**arget charact**E**rization using **NE**twork **T**opology), a network-based approach that characterizes known targets in signaling networks using topological features. Specifically, we use a SVM-based approach to identify the set of topological features (referred to as *predictive topological features*) that characterizes known targets and to generate a *characterization model* using these features. The model specifies which topological features are important for discriminating the tar-

gets and how these features should be combined to produce a quantitative *score* that identifies the likelihood of a node being a target. In particular, TENET uses *feature selection* to select *predictive topological features* and *weighted misclassification cost* to handle SVM training issues such as noisy labels and imbalanced data. Our empirical study on four real-world curated signaling networks demonstrates the effectiveness and superiority of TENET.

The rest of the paper is organized as follows. In Section **??**, we define some terminology and introduce the topological features being considered and the target curation process that is used for identifying known targets used subsequently for validation. Then, we formally define the target characterization problem and describe the TENET algorithm in Section 3. Finally, we present the experimental results in Section 4 and conclude the paper in Section 5.

2 Preliminaries

In this section, define some terminology and topological features, and introduce the target curation process that we shall be using in the sequel.

2.1 Terminology

A biological signaling network can be modelled as a directed hypergraph $G = (V, E)$ [50] where the nodes V represent molecules (*e.g.*, proteins) and the *hyperedges* E represent biochemical reactions and processes. A hyperedge connects one node set U to another W, where $U, W \subseteq V$. For instance, in the activation of ERK, the set U in the hyperedge consists of ERK and its kinase, phosphorylated MEK whereas W contains the phosphorylated ERK (ERKPP). Analysis of directed hypergraphs is generally more complex than graphs and many graph algorithms cannot be used directly on hypergraphs. Hence, they are often transformed into graphs containing simple edges for analysis. Methods (*e.g.*, bipartite and substrate graph representation) exist for such transformation [50]. In this paper, we use the bipartite graph representation as it retains the original information of the hypergraph [50]. Signaling networks generally contain characteristics such as feedback and feedforward loops which are common in complex regulatory control [56]. These loops in turn give rise to graph characteristics such as strongly connected components (SCC).

The activity of nodes in the signaling network are generally governed by complex interconnectivity of various nodes in the same network. We refer to a node as a *candidate target* if when perturbed, it modulates the activity of a specific node (referred to as *disease node*). A *disease node* is a protein that is either involved in some biological processes which may be deregulated, resulting in manifestation of a disease, or be of interest due to its potential role in the disease. For instance, in the MAPK-PI3K network [36] that is often implicated in cancer, ERKPP can be considered as a disease node due to its role in proliferation. Given a signaling

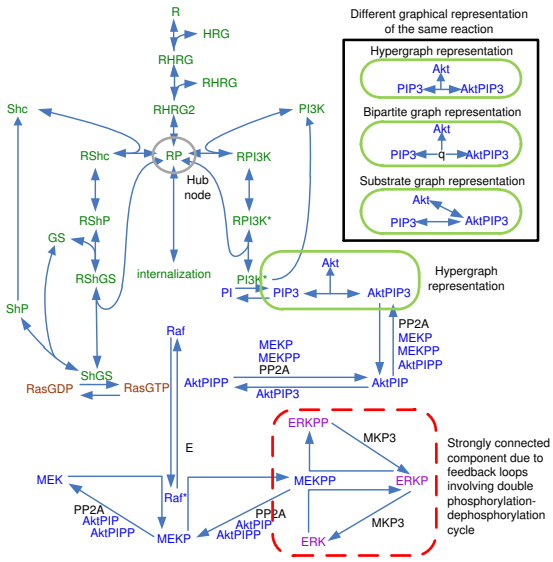

Figure 1: MAPK-PI3K network adapted from [36].

network $G = (V, E)$ and a disease node $x \in V$, let the set of nodes having a path leading to x be denoted as $V_x \subseteq V$. Then, the set of *candidate target* nodes in G relevant to x is denoted as $T_x \subseteq V_x$.

Network-based analysis can be applied to signaling networks to study the characteristics and properties of these networks. In this paper, we examine a total of sixteen topological features that are summarized in Table 1. These features are selected based on their role in measuring relative importance of a node in a signaling network. The formal definitions as well as motivation for selecting these features are given in [16] (also detailed in Section 2.2).

2.2 Topological Features

In this subsection, we introduce the 16 topological features that we study. These features are selected based on their role in measuring relative importance of a node in a signaling network. The MAPK-PI3K network (Figure 1) is used as a running example to illustrate the features.

Degree Centrality. It is a local centrality measure based on the number of edges a node has [32]. For directed networks such as the signaling network, there are three variants of *degree centrality*, namely, *in degree*, *out degree* and *total degree centrality* which consider only in-going edges, only out-going edges, and all edges of a node, respectively.

Definition 1 *Given a signaling network $G = (V, E)$,* ***in degree centrality*** *of a node $u \in V$ is defined as $\theta_{in(u)} = \sum_{v \in V} |e_{vu}|$ where $e_{vu} \in E$ is the edge connecting node $v \in V$ to u.* ***Out degree centrality*** *and* ***total degree centrality*** *are denoted*

	MAPK-PI3K (I_1)	Glucose-Stimulated Insulin Secretion (I_2)	Endomesoderm Gene Regulation (I_3)	Glucose Metabolism (I_4)
BioModel ID	BIOMD0000000146	BIOMD0000000239	BIOMD0000000235	BIOMD0000000244
Related disease or biological process	Ovarian cancer	Type 2 diabetes mellitus	Embryonic development	Glucose to acetate metabolism
No. of targets	9	6	206	16
Repository used for curation	ClinicalTrials.gov	ClinicalTrials.gov	PubMed	PubMed
Keywords used for curation	ovarian cancer drug	type 2 diabetes mellitus drug	sea urchin endomesoderm	E Coli glucose metabolism to acetate
Date of Curation	29 Apr 2014	25 Jan 2013	28 Oct 2013	14 Nov 2013
Unique Drugs Curated	458	617	-	-
Relevant Drugs Curated	22	16	-	-

Table 2: Summary of the curation results.

as $\theta_{out(u)} = \sum_{v \in V} |e_{uv}|$ *and* $\theta_{total(u)} = \theta_{in(u)} + \theta_{out(u)}$*, respectively.*

Generally, a node with high *degree centrality* (hub) is considered an important node. In particular, studies have found that biological networks resemble *scale-free networks* [75] in that they are robust against random perturbation of non-hub nodes [1]. Specifically, a high *in degree* node acts as a signal integrator by integrating multiple signals while a high *out degree* node acts as a signal differentiator. For instance, double phosphorylated MEK (MEKPP) is an out degree hub and functions as a signal differentiator.

Eigenvector Centrality. Nodes with high *eigenvector centrality* are well-connected to other central nodes [5]. In a signaling network, these nodes tend to be located in the network where signals either converge or diverge depending on whether these central nodes have high in-degree or out-degree. For instance, activated ErbB4 receptor (RP) which has high *eigenvector centrality* is connected to many other central nodes such as PI3K*, and provides a means for converging and diverging the various signals passing through the network.

Definition 2 *Given a signaling network* $G = (V, E)$*, let* N_u *be the set of neighbors of node* $u \in V$*. Then, the* ***eigenvector centrality*** *of* u *is defined as* $\alpha_u = \frac{1}{\lambda} \sum_{v \in N_u} \alpha_v$ *where* λ *is a constant.*

According to the Perron–Frobenius theorem, in the above definition λ has to be the largest eigenvalue of the adjacency matrix[2] A if the centralities are to be non-negative [71].

Closeness Centrality, Eccentricity Centrality and Proximity Prestige. These features are based on the proximity of a node to other nodes in the network. *Closeness centrality* assigns node centrality value using the sum of the shortest path distance [32] while *eccentricity centrality* uses the largest shortest path distance [103]. In contrast to *closeness centrality* which uses the set of nodes that a node u can reach (influence range), *proximity prestige* assesses importance based on the set of nodes that can reach u (influence domain).

[2]The adjacency matrix $A = \{a_{ij}\}$ specifies the connectivity of the network such that $a_{ij} = 1$ implies an edge connecting node i to j.

Definition 3 *Given a signaling network $G = (V, E)$, let $I_u \subseteq V$ be the set of nodes having at least one path leading to node u and l_{uv} be the shortest path length between nodes u and v, where $u, v \in V$. Then, the* ***closeness centrality*** *β_u,* ***eccentricity centrality*** *γ_u and* ***proximity prestige*** *μ_u of node u are defined as $\beta_u = \frac{|V|}{\sum_{v \in V} l_{uv}}$, $\gamma_u = \frac{1}{max\{l_{uv}\}}$, and $\mu_u = \frac{\frac{|I_u|}{|V|-1}}{\frac{\sum_{v \in V} l_{vu}}{|I_u|}}$, respectively.*

In a signaling network, the above measures of a node can be used to determine how central it is to the regulation of other nodes in the network [86]. For instance, ShGs which lies near the center of the network is well connected to many other nodes in the network. Hence, it has higher *closeness centrality* compared to other nodes (*e.g.*, MKP3) that lie near the boundary of the network. Also, nodes with high *eccentricity centrality* are likely to be influential signal transmitters, regulating many other nodes [86]. For instance, PI3K* which lies near the center of the network has higher eccentricity centrality compared to other fringe nodes such as ERK since the fringe nodes tend to be further away from other nodes in the network.

Betweenness Centrality. This feature assigns node centrality value based on the ease in which a node can reach other nodes in the network [6].

Definition 4 *Given a signaling network $G = (V, E)$, let $d_{st}(v)$ be the number of shortest paths from nodes s to t passing through v where $s, t, v \in V$. Then, the* ***betweenness centrality*** *of v is defined as $\delta_v = \sum_{s \neq v \neq t \in V} \frac{d_{st}(v)}{d_{st}}$.*

In a signaling network, these nodes can be considered efficient and crucial signal transmitters as they tend to lie on a majority of the shortest paths between node pairs in the network. For instance, AktPIP3, a hub node, has high betweenness centrality in the network as it is well connected to many other central nodes, hence providing fast access to other nodes in the network. Comparatively, nodes (*e.g.*, MKP3) that lie on the fringe of the network has low betweenness centrality.

Bridging Centrality and Bridging Coefficient. The *bridging centrality* identifies *bridging nodes* (nodes with high *bridging centrality* value) which are located between functional modules in the signaling network and mediate signal flow between the modules [41]. The *bridging coefficient* measures the average probability of a node transmitting signals to its direct neighbourhood.

Definition 5 *Given a signaling network $G = (V, E)$, let $\theta_{total(v)}$ be the total degree of node $v \in V$, N_v be the set of neighbors of v, and η_i be the number of outgoing edges of node i, where $i \in N_v$. Then, the* ***bridging coefficient*** *of a node v is defined as $\pi_v = \frac{1}{\theta_{total(v)}} \sum_{i \in N_v, \theta_{total(i)} > 1} \frac{\eta_i}{\theta_{total(i)} - 1}$.*

Definition 6 *Given the inverses of betweenness centrality rank and bridging coefficient rank of node v denoted as $\psi_{\frac{1}{\delta:v}}$ and $\psi_{\frac{1}{\pi:v}}$, respectively, the* ***bridging centrality*** *is defined as $\zeta_v = \psi_{\frac{1}{\delta:v}} \times \psi_{\frac{1}{\pi:v}}$.*

Ovarian Cancer Drugs in [73]	Mechanism of Action	Target in I_1
Lapatinib (Phase I) [82]	Bind to ATP-binding site of receptor (R), preventing its autophosphorylation	RP
Sorafenib (Phase II) [100]	Bind to ATP-binding site of Raf, preventing activation of Raf	Raf*
ISIS 5132 (Phase II) [18]	Bind to Raf mRNA to downregulate Raf expression	Raf
AZD6244 (Phase II) [107]	Bind and lock MEK into inactive conformation	MEKPP
XL147 (Phase I) [72]	Bind to ATP-binding site of PI3K, preventing activation of PI3K	PI3K*
Perifosine (Phase I) [54]	Bind to lipid-binding PH domain of Akt	AktPIP, AktPIPP, AktPIP3
ECO-4601 (Phase I) [7, 72]	Degrade Raf1 through proteasomal-dependent mechanism	Raf
PKI-587 (Phase I) [96]	Inhibits PI3K and MTOR kinases	PI3K*
PKI-179 (Phase I) [72]	Small-molecule mimetic of ATP that inhibits PI3K and MTOR kinases	PI3K*
BKM120 (Phase I) [72]	ATP competitive inhibitor of PI3K kinase	PI3K*
AZD5363 (Phase I) [2, 72]	ATP-competitive pan-Akt inhibitor	Akt
BYL719 (Phase I, II) [72]	Specifically inhibits PI3K in the PI3K/Akt kinase signaling pathway, thereby inhibiting the activation of the PI3K signaling pathway	PI3K*
Dabrafenib (Phase I, II) [72]	Selectively binds to and inhibits the activity of B-raf, which may inhibit the proliferation of tumor cells which contain a mutated BRAF gene	Raf*
GSK1120212 (Phase I, II) [42]	Potent and selective allosteric inhibitor of MEK1/2	MEKPP
GSK2110183 (Phase I, II) [88]	ATP-competitive pan-Akt inhibitor	Akt
GSK2141795 (Phase I) [2]	ATP-competitive Akt inhibitor	Akt
MEK162 (Phase I) [72]	Non-competitive with ATP. Binds to and inhibits the activity of MEK1/2	MEKPP
MK-2206 (Phase II) [106]	Oral pan-Akt inhibitor	Akt
Pimasertib (Phase II) [72]	Selectively binds to and inhibits the activity of MEK1/2, preventing the activation of MEK1/2-dependent effector proteins and transcription factors.	MEKPP
SAR245409 (Phase II) [72]	Inhibits both PI3K kinase and MTOR kinase	PI3K*
Trametinib (Phase II) [72]	Specifically binds to and inhibits MEK 1 and 2	MEKPP
Triciribine (Phase I, II) [72]	Inhibits the phosphorylation, activation, and signalling of Akt-1, -2, and -3	Akt

Table 3: Ovarian cancer drugs in [73] and their targets in the MAPK-PI3K network [36].

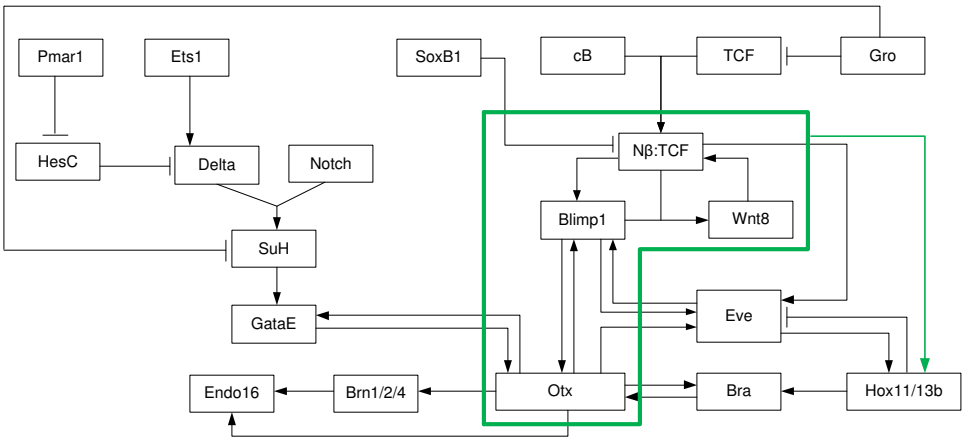

Figure 2: The sea urchin Endo16 regulatory pathway. The nodes in this pathway are targets for regulating Endo16 in I_3.

For instance, PIP3 has *high bridging coefficient* and *bridging centrality* since it is positioned at the boundary of a strongly connected component (SCC) within the network and helps to transmits signal between nodes outside the SCC and those within it.

Clustering Coefficient. This feature determines how well the neighbourhood of a node is connected [99] by considering how close the neighbourhood is to being a

T2DM Drugs/Food Constituents in [73]	Mechanism of Action	Target in I_2
Alcohol (Phase II) [22]	Causes hyperglycemia	glucose
Avandamet (Rosiglitazone + Metformin) (Phase IV) [83]	Rosiglitazone is a highly selective and potent agonist for the PPARγ. Metformin decreases hepatic glucose production, decreases intestinal absorption of glucose and increases peripheral glucose uptake and utilization	glucose
Metformin (Phase IV) [51, 98]	Inhibit respiratory-chain complex 1 of mitochondria, thereby reducing cellular energy transiently and activating AMPK. Increase peripheral glucose uptake and utilization	glucose
Benfluorex (Phase II) [53]	Reduce β-oxidation (process by which fatty acid molecules break down in mitochondria to generate acetyl-CoA) rates and ketogenesis. Reduce gluconeogenesis from lactate/pyruvate	acetyl-CoA, glucose
Berberine (Phase III) [90]	Mimick insulin action, improve insulin action by activating AMPK, reduce insulin resistance through PKC-dependent up-regulation of insulin receptor expression, inducing glycolysis, promoting GLP-1 secretion and modulating its release, inhibition of DPP-4	insulin, glucose
Cod (Phase II) [57]	Increase insulin-stimulated glucose uptake	intracellular glucose
Deproteinised hemoderivative of calf blood (Actovegin) (Phase III) [61]	Actovegin is composed of small molecules present under normal physiological conditions, therefore pharmacokinetic and pharmacodynamic studies to determine its active substance are not feasible. Increase glucose uptake and improve oxygen uptake under conditions of ischemia	glucose
Tagatose (Phase III) [25]	Interfere with carbohydrate absorption by inhibiting intestinal disaccharidases and glucose transport	plasma glucose
Vinegar [45]	Reduce plasma renin activity and aldosterone concentration. Reduce serum glucose	serum glucose
Rice [40]	Increase dietary glycemic load	blood glucose
CS-917 (Phase II) [24]	Competitively inhibits fructose-1,6-bisphosphatase (FBPase) at the AMP binding site	activated FBPase
MB07803 (Phase II) [95]	Fructose-1,6-bisphophatase (FBPase) inhibitor	activated FBPase
Glycerol [59]	Synthesized from glycerol using glycerol kinase	glycerol-3- phosphate
Methylcobalamin [30]	Provides methyl group that couples to CO to synthesize acetyl-CoA	acetyl-CoA
Gynostemma Pentaphyllum tea (Phase II) [4, 67]	Increase superoxide dismutase (SOD). Strong inhibition on IL-6 and Ptgs2 MRNA expression and weak inhibition on TNF-α MRNA expression. Gypenosides, the major components of Gynostemma pentaphyllum, increase Bax, reduce Bcl-2 and stimulate release of cytochrome c, AIF (apoptosis-inducing factor), and Endo G (endonuclease G) from mitochondria. Ferricytochrome c reduction to ferrocytochrome c may be required for depolarization of the mitochondrial	ferrocytochrome c
Xanthohumol [4, 19]	Binds annexin V-FITC, cleaves PARP-1 and activates procaspases-3, -8 and -9. Depolarizes mitochondrial leading to release of cytochrome c. Inhibits Akt, NFkB, MTOR, Bcl-2 and survivin. Ferricytochrome c reduction to ferrocytochrome c may be required for depolarization of the mitochondrial	ferrocytochrome c

Table 4: T2DM drugs or food constituents in [73] and their targets in the glucose-stimulated insulin secretion network [44].

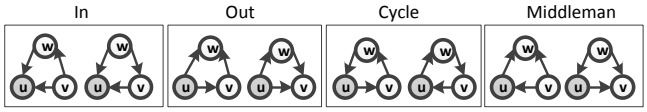

Figure 3: Directed triangle graphs adapted from [26].

clique where every node within the clique is connected to every other node in it [1]. The original definition was meant for undirected graph. A variety of definition exists [26] when edge directions are considered (Figure 3).

Definition 7 *Given a signaling network $G = (V, E)$, let $e_{ij} \in E$ denote an edge connecting nodes i to j where $i, j \in V$ and $A = \{a_{ij}\}$ be the adjacency matrix where $a_{ij} = 1$ if and only if $\exists e \in \{e_{ij}, e_{ji}\} \subseteq E$ and zero otherwise. Then, the **undirected-**, **in-**, **out-**, **cycle-** and **middleman-clustering coefficient** of a node $u \in V$ denoted as $\kappa_{undir(u)}$, $\kappa_{in(u)}$, $\kappa_{out(u)}$, $\kappa_{cyc(u)}$ and $\kappa_{mid(u)}$, respectively, are defined as $\kappa_{undir(u)} = \frac{(A^3)_{ii}}{\theta_{total(u)}(\theta_{total(u)}-1)}$, $\kappa_{in(u)} = \frac{(A^T A^2)_{ii}}{\theta_{in(u)}(\theta_{in(u)}-1)}$, $\kappa_{out(u)} =$*

Target ID in BIOMD0000000244 (I_4)	Name	Reference
ACT	acetate	*
GLC	glucose	*
G6P	glucose-6-phosphate	[109]
ICT	isocitrate	[94]
PEP	phosphoenolypyruvate	[10, 27]
AceB	malate synthase A	[77]
Acoa2act	enzyme for reaction from ACoA to ACT	[9]
Cya	adenylate cyclase	[76]
Fdp	fructose-1,6-bisphosphatase I	[78]
Icd	unphosphorylated isocitrate dehydrogenase	[46]
Icd_P	phosphorylated isocitrate dehydrogenase	[46]
Pdh	pyruvate dehydrogenase	[69]
Ppc	phosphoenolpyruvate carboxylase	[27]
PpsA	phosphoenolpyruvate synthase	[74]
EIIA	unphosphorylated PTS protein EIIA	[43]
EIICB	PTS protein EIICB (ptsG)	[79]

Table 5: Targets that are crucial for acetate production in the glucose metabolism network in [55]. * indicates targets that are included by default due to their direct involvement (either as input or output) in the metabolic reaction being studied.

$\frac{(A^2A^T)_{ii}}{\theta_{out(u)}(\theta_{out(u)}-1)}$, $\kappa_{cyc(u)} = \frac{(A^3)_{ii}}{\theta_{in(u)}\theta_{out(u)}-A_{ii}^2}$ *and* $\kappa_{mid(u)} = \frac{(AA^TA)_{ii}}{\theta_{in(u)}\theta_{out(u)}-A_{ii}^2}$ *where* $\theta_{in(u)}$, $\theta_{out(u)}$ *and* $\theta_{total(u)}$ *are the in, out and total degree of* u*, respectively;* A^T *is the transpose of* A*;* A^n *is the matrix product of* n *copies of* A*; and* A_{ii} *denotes the* i^{th} *element of the main diagonal of* A*.*

Note that in the above definition, the neighbourhood size must be greater than one. For smaller neighbourhood sizes ($N_u = 0$ and $N_u = 1$), the coefficients are set to zero.

Target Downstream Effect (TDE). TDE assesses the potential impact on the network when a node is perturbed based on the probability of perturbing a downstream node[3] w and the likelihood of w causing off-target effect [15].

Definition 8 *Given a signaling network* $G = (V, E)$*, let* W *be the set of downstream nodes of* $v \in V \setminus W$*. Let* $\rho_{v,w}$ *be the probability of perturbing* $w \in W$ *when target node* v *is perturbed and* $\theta_{total(w)}$ *be the total degree of* w*. The* ***target downstream effect*** *of* v *is defined as* $\omega_v = \sum_{w \in W}(\rho_{v,w} \times \theta_{total(w)})$*.*

2.3 Target Curation Process

In this section, we describe the target curation process used to identify the set of benchmark targets required by TENET. Manual curation of literature generates substantially lower error rates than text mining-based approaches [89]. In this article, we study two categories of networks: networks associated with human diseases

[3]Node w is *downstream* of v if there exists a path from v to w.

and networks describing biological processes. Table 2 summarizes the curation results of the four signaling networks we studied.

Human Disease-Related Networks. Amongst the four networks we study, two are associated with human diseases. These networks are the `MAPK-PI3K` [36] and the `glucose`-stimulated `insulin` secretion networks [44]. The curation process for these networks is as follows:

1. Obtain a list of unique drugs and compounds relevant to the human disease from clinical trial database [73].
2. Obtain the targets of these drugs and compounds via drug related databases [102] and literature survey.
3. Identify the targets that are in the scope of the signaling network.

Biological Process-Related Networks. The remaining networks we studied describe specific biological processes of particular organisms. The curation process for these networks is as follows:

1. Obtain a list of unique molecules (genes or proteins) relevant to the biological process of the specific organism from *PubMed* using specific keywords.
2. Identify the molecules that are in the scope of the signaling network.

The curated targets are listed in Tables 3 to 5 and Figure 2.

3 Target Characterisation

In this section, we formally define the target characterization problem and describe the TENET algorithm.

3.1 Topological Feature-based Target Characterization

Intuitively, the goal of topological feature-based target characterization is to use a set of *predictive topological features* to characterize known targets in a network. Hence, the *topological feature-based target characterization problem* can be formulated as a supervised learning problem. In a supervised learning problem, a training set $\{\langle x_i, f(x_i)\rangle\}$ is given where $f(x_i)$ is the predictor of x_i and the goal is to learn some target function $f : X \rightarrow Y$ which can be applied to predict unseen data w. The problem can be subdivided into two categories: regression when the predictor yields a continuous outcome and classification when the outcome is discrete. A regression problem can be converted into a binary classification problem by specifying a threshold h and assigning x_i with f greater than h to one class and the remaining to the other class. We advocate that the *topological feature-based target characterization problem* is best represented as a regression problem. In this problem, we are interested in finding out how likely one node is a target relative to another node based on a set of predictive topological features. This is

different from the target classification problem where we want to find out the class membership of a node. Note that the regression problem can be converted into a classification problem by specifying a threshold h and assigning nodes having target function greater than h to the target class and the rest to the non-target class.

Although we examine sixteen topological features, as we shall see later, not all features are relevant to a given signaling network. In fact, incorporating irrelevant features may adversely impact the performance of the prediction model. Hence, it is important to learn a set of predictive topological features that best characterizes targets (referred to as *topological feature selection*) for a given network. Formally, it is defined as follows.

Definition 9 Given a signaling network $G = (V, E)$ and a disease node $x \in V$, let $T_x \subseteq V$ and $\mathcal{X}_{all}$ denote the set of known targets in G relevant to x, and the set of topological features of G, respectively. Then, the goal of **topological feature selection** is to find a set of **predictive topological features** $\mathcal{F} \subseteq \mathcal{X}_{all}$ that maximizes the prediction accuracy for $f(\xi(u, \mathcal{F}))$ subject to the following conditions:

$$\begin{cases} f(\xi(u, \mathcal{F})) = 1 & \text{when } u \in T_x, \\ f(\xi(u, \mathcal{F})) = 0 & \text{otherwise}. \end{cases} \tag{1}$$

Then the *topological feature-based target characterization problem* is formally defined as follows.

Definition 10 Given a signaling network $G = (V, E)$, a disease node $x \in V$, T_x, and $\mathcal{X}_{all}$, let $\mathcal{F}$ denote the set of predictive topological features. Then, for a threshold h, the goal of the **topological feature-based target characterization problem** is to identify a set of predictive topological features $\mathcal{F} \subseteq \mathcal{X}_{all}$ using topological feature selection and learn a **characterization model** $g(\xi(u, \mathcal{F}))$ subject to the conditions

$$\begin{cases} g(\xi(u, \mathcal{F})) \in \Re, & \\ g(\xi(u, \mathcal{F})) \geq h & \text{when } u \in T_x, \\ g(\xi(u, \mathcal{F})) < h & \text{otherwise}, \end{cases} \tag{2}$$

that maximizes the target prediction for $g(\xi(u, \mathcal{F}))$.

Figure 4 depicts a pictorial overview of the topological feature-based target characterization problem. For example, given the MAPK-PI3K signaling network, its associated disease node ERKPP, the set of known targets in this network, and the topological features in Table 1, the goal of this problem is to produce the followings: (a) Identify the set of predictive topological features $\mathcal{F} = \{\delta, \pi, \theta_{in}, \theta_{out}\}$ and (b) learn a characterization model $g(\xi(\texttt{ERKPP}, \mathcal{F}))$. Note that in Definition 10, there is no need to explicitly specify a threshold h if we are only interested in obtaining the relative rankings of the nodes. The threshold is required if we want to assign class labels (*e.g.*, target class) to the nodes.

3.2 SVM-based Target Characterization

We employ *support vector classification* (SVC) to select predictive topological features and *support vector regression* (SVR) to generate the characterization model. The SVC and SVR are typically formulated as constrained optimization problems and solved using the *Lagrangian multiplier method*. In general, SVM models contain multiple parameters, such as the cost parameter C and parameters related to the kernel function, that affect the learning and performance of the models [11]. We follow the method in [39] for training the SVM. The feature values are scaled linearly to the range of [0, 1] for each signaling network to avoid features with larger ranges dominating those with smaller ranges. We use stratified[4] cross-validation (described below) and grid-search [39] on the training data to identify the best values of the model parameters. Note that cross-validation helps us to avoid the issue of overfitting the data whereas stratification enables us to keep the percentage of targets in the different folds similar to the original dataset. The best parameter is the one that yields the best average prediction accuracy for the cross-validation process. Wherever possible[5], we use a 10-fold stratified cross-validation since larger fold numbers reduce pessimistic bias and 10 folds generally give good performances [52].

Several non-trivial issues, namely, irrelevant or redundant features, noisy labels and imbalanced data set, need to be addressed in training the SVM model for characterizing targets. In particular, we use feature selection to select for appropriate features to be used in the SVM model and cost-sensitive learning to handle the issue of noisy labels and imbalanced data set. We examine three feature selection approaches, namely, backward stepwise elimination (BSE) [63], Wilcoxon-ROC based elimination (WRE) and WRE-BSE. BSE is *classifier-aware* whereas WRE is *classifier-independent*. WRE-BSE which performs WRE followed by BSE is a hybrid approach. Note that compared to classifier-independent methods, classifier-aware methods interact with the classifiers and such interaction can lead to better classification results [84]. However, they are typically computationally expensive and run the risk of model-overfitting. Cost sensitive learning is an algorithmic approach that chooses an appropriate strategy specific to the classifier to overcome the bias introduced by imbalanced data and the noise caused by uncertainty in labelling. We use *weighted misclassification cost* (WMC), an approach that proportionates the misclassification cost of the training data according to class. In particular, we use a variable C_i as the cost parameter C:

$$C_i = \begin{cases} C^+ & \text{if } y_i = +1 \\ C^- & \text{if } y_i = -1 \end{cases} \tag{3}$$

subject to the constraints $C^+ + C^- = 1$, $C^+ > 0$ and $C^- > 0$ where y_i is the

[4]The training data was sampled from the original data such that the ratio of the targets to non-targets is similar to that of the original data.

[5]In our study, we set a lower bound of one target in all our test sets.

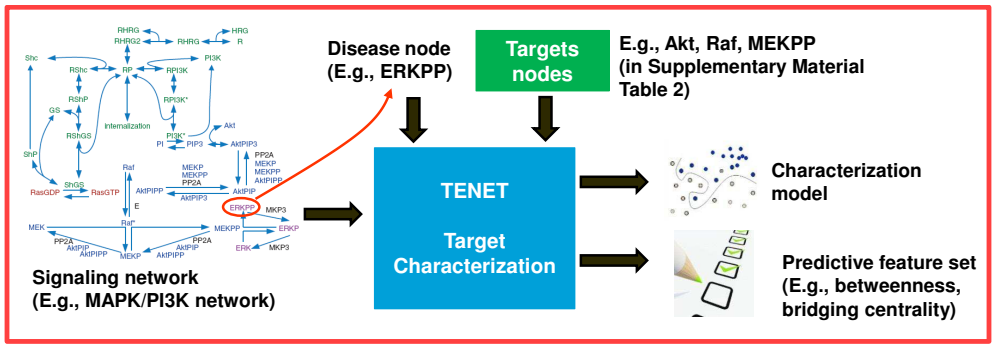

Figure 4: Target characterization problem.

Network ID	Test Set
I_1	MEKPP$^\sharp$, RPI3K, internalization, PP2A
I_2	FBP$^\sharp$, OXA_{cyt}, $Malate_{cyt}$, Aspartate, Glutamate, Citrate, ATP, $NADP^+$, NAD^+, Ubiquinone
I_3	GENE_E_Otx$^\sharp$, GENE_E_Pmar1$^\sharp$, GENE_M_Eve$^\sharp$, GENE_P_Bra$^\sharp$, GENE_P_Eve$^\sharp$, PRE_E_UbiqSoxB1$^\sharp$, PRE_P_cB$^\sharp$, PROTEIN_E_HesC$^\sharp$, PROTEIN_E_UVAotx$^\sharp$, PROTEIN_E_cB$^\sharp$, PROTEIN_E_nBTCF$^\sharp$, PROTEIN_M_Pmar1$^\sharp$, PROTEIN_M_SoxB1$^\sharp$, PROTEIN_P_Delta2$^\sharp$, PROTEIN_P_SoxB1$^\sharp$, mRNA_E_Ets1$^\sharp$, mRNA_M_HesC$^\sharp$, mRNA_M_SoxB1$^\sharp$, mRNA_M_UbiqSoxB1$^\sharp$, mRNA_P_Endo16$^\sharp$, mRNA_P_UbiqEts1$^\sharp$, GENE_E_Apobec, GENE_Ees, GENE_E_Kakapo, GENE_E_OrCt, GENE_E_Sm50, GENE_E_SuTx, GENE_M_Alx1, GENE_Mcapk, GENE_M_FoxO, GENE_M_Nrl, GENE_M_VEGFR, GENE_P_OrCt, GENE_P_Sm27, PRE_E_VEGF, PROTEIN_E_Apobec, PROTEIN_E_Ficolin, PROTEIN_E_Hex, PROTEIN_E_OrCt, PROTEIN_E_Sm27, PROTEIN_M_FoxN23, PROTEIN_M_L1, PROTEIN_M_Sm27, PROTEIN_M_UbiqSoxC, PROTEIN_P_Dpt, PROTEIN_P_Dri, PROTEIN_P_FoxA, PROTEIN_P_Pks, mRNA_E_Dri, mRNA_E_ES, mRNA_E_Gcm, mRNA_M_Gcad, mRNA_M_Hex, mRNA_M_Not, mRNA_M_Snail, mRNA_M_TBR, mRNA_M_Tel, mRNA_M_z13, mRNA_P_Apobec, mRNA_P_Dri, mRNA_P_OrCt
I_4	Isocitrate$^\sharp$, EIIA$^\sharp$, Emp, Enolase, Crp

Table 6: Test set of I_1 to I_4. Nodes marked with $^\sharp$ are known targets.

class predictor and C^+ and C^- denote the misclassification cost of the target and non-target classes, respectively.

Data partitioning (stratified sampling). Note that for the networks studied, the target class for all nodes including the test set is known. This is for the purpose of validating our approach later in the experiments. We partition the data into training and test set by following two rules. First, there should be at least one target in the test set. This allows us to determine if TENET is able to rank the curated target higher than other nodes. Second, the ratio of target nodes to non-target nodes should mirror that of the original data set as close as possible. This ensures the real distribution of targets versus non-targets in the networks is retained. Using these two rules, we determine the number of targets and non-targets in the test set for each network. Then, the targets and non-targets are randomly selected from the original data set to generate the test set. Finally, the remaining nodes form the training set. The test set of I_1 to I_4 are provided in Table 6. Note that these same rules are followed when generating individual folds from the training set.

Cross validation. We use cross validation (illustrated in Figure 5) for training the SVM. Briefly, the training data (two matrices[6], one for candidate targets and one for candidate non-targets) is partitioned using stratified sampling into multiple

[6]The rows represent nodes and columns represent topological features

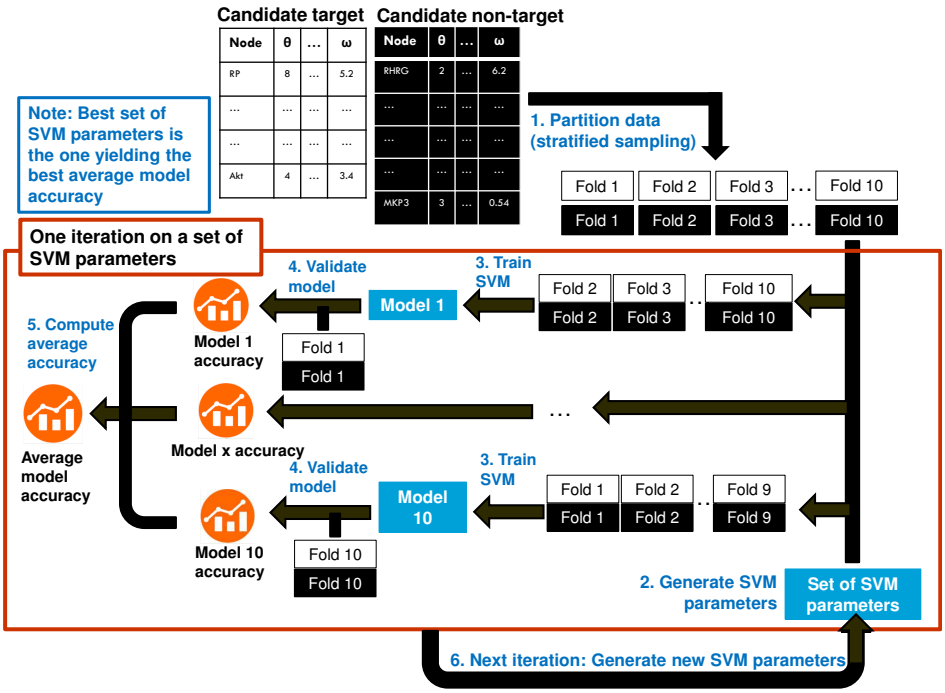

Figure 5: Example of a 10-fold cross validation (best viewed in colour).

folds. For the training, multiple iterations are needed to explore different sets of SVM parameters. In each iteration, x models corresponding to the number of folds are generated. For each model, one of the folds is excluded[7] from the training and used for validating the model. Then, the model accuracies are averaged. Cross validation terminates when the SVM parameters have been satisfactorily explored. In TENET, we use the grid-search approach (detailed in [39]) for exploring the SVM parameters. The best SVM model is the one with parameters yielding the highest average model accuracy.

3.3 The TENET algorithm

Given a signaling network $G = (V, E)$, a disease node $x \in V$, a known target set $T_x \subseteq V$, a set of topological features $\mathcal{X}_{all}$ and a step-size of the misclassification cost s, TENET identifies the set of predictive structural features and a characterization model that best characterizes these known targets. Note that $\mathcal{X}_{all}$ and s are optional inputs and are set to default values[8] if they are not given. The known targets T_x can be extracted by following the curation process described in [16] (Section 2.3).The TENET algorithm is given in Algorithms 1 to 5 and is comprised of three phases, namely, the *pruning phase* (Algorithm 2), the *feature extraction phase* and the *model training phase*. First, the *pruning phase* identifies relevant

[7] Note that each fold is being excluded only once from the SVM training.

[8] $\mathcal{X}_{all}$ is set to the 16 topological features given in Table 1 whereas s is set to 0.1.

Algorithm 1: Algorithm TENET

Input: Signaling network G, disease node x, target set T_x, topological feature set $\mathcal{X}_{all}$ (optional) and step-size of the misclassification cost s (optional).
Output: Predictive topological feature set $\mathcal{F}$ and characterization model $\mathcal{M}$.

$\mathcal{F}, \mathcal{M}, \mathcal{X}_{all} \leftarrow$INITIALIZE($\mathcal{F}, \mathcal{M}, \mathcal{X}_{all}$)
$V_{candidate} \leftarrow$FILTERCANDIDATE(G, x); // Phase 1
$H \leftarrow$EXTRACTFEATURE($G, V_{candidate}, \mathcal{X}_{all}$); // Phase 2
$\mathcal{F}, \mathcal{M} \leftarrow$TRAINMODEL($H, T_x, s$); // Phase 3

Algorithm 2: Procedure FILTERCANDIDATE

Input: Signaling network G, disease node x.
Output: Candidate target node set $V_{candidate}$.

$G_{\text{BI}} \leftarrow$CONVERT2BIPARTITEGRAPH(G)
$G_{\text{DAG}} \leftarrow$CONVERT2DAG($G_{\text{BI}}$)
$U \leftarrow$GETROOTNODES(G_{DAG})
foreach *iteration* i*=1 to* $|U|$ **do**
    $G_{\text{DAG}} \leftarrow$INDEX($G_{\text{DAG}}, U_i$, null)
$V_{candidate} \leftarrow$ASSESSREACHABILITY(G_{DAG}, x)

nodes (denoted as $V_{candidate}$) that shall be used for training the SVM. Then, the *feature extraction phase* extracts all the topological features (denoted as $\mathcal{X}_{all}$) of each candidate node and stores them in a $|V_{candidate}| \times |\mathcal{X}_{all}|$ matrix H. Finally, in the *model training phase*, TENET learns the optimal set of predictive topological features $\mathcal{F}$ and the best model parameters of the characterization model $\mathcal{M}$. We shall now describe these phases in turn.

Phase 1: Pruning. In this phase, TENET prunes nodes that do not have paths leading to the disease node x. This phase yields a set of potential candidate nodes $V_{candidate} \subset V$ and is used to reduce the subsequent computation. In the pruning process, the given network G is first preprocessed into a bipartite graph and then converted into a *directed acyclic graph* (DAG) (Algorithm 3), a graph with consistent topological ordering, to facilitate indexing of nodes (Algorithm 4). Note that the node indices shall be used subsequently to perform reachability check to identify the nodes to be pruned. We adopt the method in [23] for bipartite graph conversion. In order to convert the bipartite graph into its DAG representation, we adopt the approach in [92] to identify SCCs and replace each SCC with a representative node (referred to as *meta node*). Then, we adopt the indexing approach of [13] to index the DAG. This indexing approach performs depth-first traversal to assign each node v a *preorder index* (when v is first visited) and a *postorder index* (when all descendent nodes of v are visited). Finally, an index-based reachability algorithm is used to determine if there exists a path from each node v to the disease node x (denoted as $v \longrightarrow x$). Given a node v and x, let w be the descendent of v that is not in the *spanning tree* (referred to as *non-spanning tree node*) and $v.preorder$ and $v.postorder$ denote the preorder and postorder indexes of v, respectively. A path $v \longrightarrow x$ exists if any of the following conditions are satisfied [13]:

1. $v.preorder \leq x.preorder$ and $v.postorder \geq x.postorder$

Algorithm 3: Procedure CONVERT2DAG

Input: Bipartite graph $G_{\text{BI}} = (V_{\text{BI}}, E_{\text{BI}})$.
Output: Directed acyclic graph (DAG) G_{DAG}.

$G_{\text{DAG}} \leftarrow G_{\text{BI}}$
$G_{\text{DAG}}.\text{SCC} \leftarrow$GETSCC($G_{\text{DAG}}$)
foreach *iteration* i*=1 to* $|G_{\text{DAG}}.\text{SCC}|$ **do**
  $V \leftarrow$INSERTNODE($V, v_{meta:i}$)
  $X \leftarrow$GETNODESINSCC($G_{\text{DAG}}.\text{SCC}_i$)
  foreach *iteration* j*=1 to* $|X|$ **do**
    $N \leftarrow$GETNEIGHBOURSNOTINSCC(X_j, X)
    foreach *iteration* k*=1 to* $|N|$ **do**
      $E \leftarrow$REPLACEEDGE($E, (X_j, N_k), (v_{meta:i}, N_k)$)
      $E \leftarrow$REPLACEEDGE($E, (N_k, X_j), (N_k, v_{meta:i})$)
      $V \leftarrow$REMOVENODE(V, X_j)

Algorithm 4: Procedure INDEX

Input: DAG $G_{\text{DAG}} = (V_{\text{DAG}}, E_{\text{DAG}})$, child node u, parent node v.
Output: DAG G_{DAG}.

if $v.preorder$*=null* **then**
  $v.preorder \leftarrow$SETTONEXTINDEX($v.preorder$)
  foreach *iteration* i*=1 to* $|V_{\text{DAG}}|$ **do**
    $w \leftarrow$GETCHILDNODE(v)
    $G_{\text{DAG}} \leftarrow$INDEX($G_{\text{DAG}}, w, v$)
    $v.descendents \leftarrow$INSERTNODE($v.descendents, w$)
    $v.descendents \leftarrow v.descendents \bigcup w.descendents$
    $v.\text{NST}Nodes \leftarrow v.\text{NST}Nodes \bigcup w.\text{NST}Nodes$
  $v.postorder \leftarrow$SETTONEXTINDEX($v.postorder$)
else if $v.preorder \neq$ *null and* $u \neq$ *null* **then**
  $u.descendants \leftarrow$INSERTNODE($u.descendents, v$)
  $u.\text{NST}Nodes \leftarrow$INSERTNODE($u.\text{NST}Nodes, v$)
if $u \neq$ *null* **then**
  $u.descendents \leftarrow u.descendents \bigcup v.descendents$
  $u.\text{NST}Nodes \leftarrow u.\text{NST}Nodes \bigcup v.\text{NST}Nodes$

2. $w.preorder \leq x.preorder$ and $w.postorder \geq x.postorder$

Note that the pruning step is beneficial in improving execution time for larger sparsely connected networks and for disease node that are positioned further upstream. For instance, in the `MAPK-PI3K` network, no nodes are pruned when we select `ERKPP` (downstream) as the disease node whereas 17 nodes (47.2%) are pruned when activated `Ras` (`RasGTP`) (upstream) is selected.

Phase 2: Feature Extraction. In this phase, for all nodes in $V_{candidate}$, TENET extracts all the topological features in Table 1 for characterizing the known targets.

Phase 3: Model Training. Given a matrix of topological feature values H, a target set T_x and a step-size of the misclassification cost s, this phase identifies a set of predictive topological features $\mathcal{F}$ and the best parameters for configuring the characterization model $\mathcal{M}$. First, the misclassification cost of the target class C^+ is initialized to a default value of 0.5. Then, feature selection is used to obtain the predictive topological feature set $\mathcal{F}$. We iterate over three different feature selection approaches (BSE, WRE and WRE-BSE). Next, the step-size s is used to step

Algorithm 5: Procedure TRAINMODEL

Input: Matrix of topological feature values H, known target set T_x and step-size of the misclassification cost s.
Output: Predictive topological feature set $\mathcal{F}$ and characterization model $\mathcal{M}$.

$C^+ \leftarrow$INITIALIZE()
$\phi_{best}, \mathcal{F}, \mathcal{M} \leftarrow$SELECTFEATURES($H, T_x, C^+$)
foreach *iteration* i=1 *to* $\frac{1}{s} - 1$ **do**
    $C^+ \leftarrow i \times s$
    $\phi_{curr} \leftarrow$TUNESVM($\mathcal{F}, T_x, C^+$)
    if $\phi_{curr} > \phi_{best}$ **then**
        $\phi_{best} \leftarrow \phi_{curr}$
        $\mathcal{M} \leftarrow$SETBESTPARAMWEIGHT($\mathcal{M}, C^+$)

Algorithm 6: Algorithm WRE

Input: Matrix of topological feature values H, target set T_x and misclassification cost for target class C^+.
Output: Prediction accuracy ϕ, predictive topological feature set $\mathcal{F}$ and characterization model $\mathcal{M}$.

$\mathcal{R}_{Wilcoxon} \leftarrow$WILCOXONFILTER($H, T_x$)
$\mathcal{R}_{\text{ROC}} \leftarrow$ROCFILTER($H, T_x$)
$\mathcal{F} \leftarrow \mathcal{R}_{Wilcoxon} \bigcap \mathcal{R}_{\text{ROC}}$
$\phi, \mathcal{M} \leftarrow$TUNESVM($\mathcal{F}, T_x, C^+$)

through the range of misclassification cost (0 – 1). In each iteration, the misclassification cost of the target class C^+ is incremented according to the number of iterations completed, before the SVM training (Algorithm 5) is performed to obtain the parameter settings of the characterization model $\mathcal{M}$ with the best accuracy.

The BSE approach is a well-known greedy approach that progressively removes features from the naïve SVM model (built using all topological features) and trains a new best model after each feature removal. The elimination process stops when removal of additional features result in a worse average accuracy of the validation set prediction. In contrast, the WRE approach (Algorithm 6) performs two statistical tests, namely, one-tailed Wilcoxon Rank-Sum (referred to as Wilcoxon) and receiver operating characteristics (referred to as ROC). The results are used to eliminate features that do not discriminate between targets and non-targets in a significant manner (based on Wilcoxon) and that do not classify targets well (based on ROC). Note that we perform two 1-tailed Wilcoxon tests and for each test; p-values smaller than 0.05 are considered significant. Hence, we take the difference of the p-values for both test hypotheses (referred to as p-value difference) and remove features with p-value difference less than 0.9. For the ROC analysis, features with AUC less than 0.7 [38] are considered poor performers and are removed. The best characterization model is found by training the SVM using the remaining features. The WRE-BSE approach (Algorithm 7) first performs WRE followed by BSE.

Algorithm 7: Algorithm WRE-BSE

Input: Matrix of topological feature values H, target set T_x, misclassification cost for target class C^+
Output: Prediction accuracy ϕ, predictive topological feature set $\mathcal{F}$, characterization model $\mathcal{M}$.

$\mathcal{R}_{Wilcoxon} \leftarrow$WILCOXONFILTER($H$,$T_x$)
$\mathcal{R}_{\text{ROC}} \leftarrow$ROCFILTER($H$,$T_x$)
$\mathcal{F} \leftarrow \mathcal{R}_{Wilcoxon} \bigcap \mathcal{R}_{\text{ROC}}$
$\phi_{prevBest}, param_{prevBest} \leftarrow$TUNESVM($\mathcal{F}, T_x, C^+$)
repeat
  foreach *iteration* i*=1 to* $|F|$ **do**
    $\phi_i, param_i \leftarrow$TUNESVM($\mathcal{F} - \mathcal{F}_i, T_x, C^+$)
  $\mathcal{F}_i, \phi_{currBest}, param_{currBest} \leftarrow$GETFEATURETOREMOVE($\phi_i, param_i$)
  if $\phi_{currBest} > \phi_{prevBest}$ **then**
    $\mathcal{F} \leftarrow \mathcal{F} \backslash \mathcal{F}_i$
    $\mathcal{M} \leftarrow param_i$
until $|\mathcal{F}| = 1$ *or* $\phi_{currBest} < \phi_{prevBest}$;

Structural Features	Time Complexity
Degree centrality	$O(\|V\|)$
Eigenvector centrality	$O(\|V\|^2)$ [49]
Closeness centrality	$O(\|V\|^3)$ [48]
Eccentricity centrality	$O(\|V\|\|E\|)$ [91]
Betweenness centrality	$O(\|V\|\|E\|)$ [6]
Bridging centrality	$O(\|V\|^2 + \|E\|)$
Bridging coefficient	$O(\|V\|^2)$
Clustering coefficient	$O(\|V\|^{2.373})$ [101]
Proximity prestige	$O(\|V\|^2 + \|E\|)$
Target downstream effect	$O(\|V\|^2 + \|E\|)$

Table 7: Time complexity for computing the different features. The proofs of those algorithm complexities that are provided without citation are given in Section 3.4.

3.4 Complexity Analysis

In this subsection, we present the complexity analysis of TENET. We start by providing the complexity analysis for the computation of the topological features considered (summarized in Table 7) in TENET.

Degree Centrality.

Theorem 1 *Computation of degree centrality requires $O(|V|)$ time in the worst case.*

Proof 1 *It takes $O(|V|)$ time to iterate through all the nodes in the graph since it requires constant time to retrieve the number of edges associated to each node.*

Bridging Coefficient.

Theorem 2 *Computation of bridging coefficient requires $O(|V|^2)$ time in the worst case.*

Proof 2 *For each node in the network, the computation of the bridging coefficient iterates through all the neighbours of the node (Definition 5). In the worst case, the*

network is a single strongly connected component where every node is a neighbour of all other nodes. Hence, calculating the bridging coefficient of all nodes in the network requires $O(|V|^2)$ in the worst case.

Bridging Centrality

Theorem 3 *Computation of bridging centrality requires $O(|V|^2)$ time in the worst case.*

Proof 3 *For each node in the network, the computation of the bridging centrality is a product of the inverse of the betweenness rank and the bridging coefficient rank (Definition 6). Hence, computation requires $O(|V|^2 + |V| + |E|)$ time in the worst case. This can be further simplified into $O(|V|^2 + |E|)$.*

Proximity Prestige

Theorem 4 *Computation of proximity prestige requires $O(|V|^2 + |E|)$ time in the worst case.*

Proof 4 *In Definition 3, the set of nodes having at least one path leading to node u (I_u) and the shortest path distance are needed for calculating the prestige value. Using the* ASSESSREACHABILITY *procedure in Algorithm 2 to obtain I_u requires $O(|V|^2 + |E|)$ time. space is required to store the node and edge information of the input graph. The shortest path distance can be found using Dijkstra's algorithm [20] and the computation requires $O(|E| + |V|log_2|V|)$ time using Fibonacci heaps [31]. In the worst case, $O(|V|^2 + |E|)$ time is needed since $O(|V|^2) > O(|V|log_2|V|)$.*

Target Downstream Effect

Theorem 5 *Computation of target downstream effect requires $O(|V|^2 + |E|)$ time in the worst case.*

Proof 5 *According to Definition 8, the computation of the target downstream effect for each node requires iterating through each of its downstream nodes. The downstream nodes can be found by using the* ASSESSREACHABILITY *procedure in Algorithm 2. The time required to compute the reachability of the nodes is $O(|V|^2 + |E|)$. In the worst case, the network is a single strongly connected component where every node has a path leading to all other nodes. Hence, calculating the target downstream effect of all nodes in the network requires $O(2|V|^2 + |E|)$ in the worst case and can be simplified to $O(|V|^2 + |E|)$.*

TENET Algorithm

Theorem 6 *Given a signaling network G, a disease node x, a target set T_x, a feature set $\mathcal{X}_{all}$ and the step-size of the misclassification cost s, the Algorithm* TENET *has worst-case time complexity $O((|V| + |E|)^2 + O(\mathcal{G}(\mathcal{X}_{all})) + O(\mathcal{T}(\cdot)))$ where $\mathcal{G}(\mathcal{X}_{all})$ is the worst-case time complexity for extracting the features and $O(\mathcal{T}(\cdot))$ is the worst-case time complexity of the feature selection method used.*

Proof 6 *In the* FILTERCANDIDATE *algorithm, the conversion of the input signaling network* $G = (V, E)$ *to a bipartite graph* $G_{\text{BI}} = (V_{\text{BI}}, E_{\text{BI}})$ *takes* $O(|V_{\text{BI}}| + |E_{\text{BI}}|)$ *time. In* DAG *conversion,* $O(|V_{\text{BI}}| + |E_{\text{BI}}|)$ *time is required for finding* SCC *using [92]. In the worst case, the signaling network is a complete directed graph and* CONVERT2DAG *takes* $O(|V_{\text{BI}}|^2 + |E_{\text{BI}}|)$ *time since* $|V_{\text{BI}}| < |V_{\text{BI}}|^2$*. In the indexing of the* DAG *graph* $G_{\text{DAG}} = (V_{\text{DAG}}, E_{\text{DAG}})$*, the depth-first traversal requires* $O(|V_{\text{DAG}}| + |E_{\text{DAG}}|)$ *time while computing the set of nodes that can reach* x *takes* $O(|V_{\text{DAG}}|)$ *time. Hence,* FILTERCANDIDATE *algorithm takes* $O(|V_{\text{BI}}|^2 + |E_{\text{BI}}|)$ *time since* $|V_{\text{BI}}| = |V| + |E|$*,* $|E_{\text{BI}}| = \sum_{(U,W)\in E} (|U| + |W|)$ *and* $(|V_{\text{BI}}| + |E_{\text{BI}}|) \geq (|V_{\text{DAG}}| + |E_{\text{DAG}}|)$*.*

The time complexity of the EXTRACTFEATURE *procedure (denoted as* $O(\mathcal{G}(\cdot))$*) depends on the features to be extracted. Amongst the features we consider, closeness centrality has the highest time complexity (*$O(|V|^3)$*) (Table 7).*

The time complexity of the TRAINMODEL *procedure is dependent on the time complexities of the feature selection approach (denoted as* $O(\mathcal{T}(\cdot))$*) and the training of the misclassification cost. Three feature selection approaches are explored. In* BSE*, a greedy approach is used for selecting a feature for removal at each iteration and a new* SVM *model is trained and tuned accordingly using the* TUNESVM *procedure.* TUNESVM *uses the grid search approach described in [8] to tune the* SVM *parameters. The tuning process takes* $O(i^p \times k)$ *where* i*,* p *and* k *are the number of iterations*[9] *required for the grid search, the number of parameters to be tuned, and the time complexity of training a* SVM*, respectively. According to [93], standard* SVM *training takes* $O(m^3)$ *time where* m *is the training set size. Hence,* TUNESVM *has* $O(i^p \times |V|^3)$ *time complexity since the training set size is approximately equal to the data set size (*$|V|$*). In the worst case, algorithm* BSE *removes all but one feature. This takes* $O(|\mathcal{X}_{all}|^2 \times i^p \times |V|^3)$ *time where* $\mathcal{X}_{all}$ *is the set of topological features. In* WRE*, the statistical-based filter requires two steps, namely,* WILCOXONFILTER *and* ROCFILTER *to find the predictive feature set. Performing the Wilcoxon test for a particular topological feature requires* $O((gh)^2)$ *time [70] where* g *and* h *are the target and non-target class sizes, respectively, and* $|V| = g+h$*. Generating* ROC *for a particular topological feature requires* $O(|V|^2)$ *time [28]. Hence,* WILCOXONFILTER *and* ROCFILTER *require* $O((gh)^2|\mathcal{X}_{all}|)$ *and* $O(|V|^2|\mathcal{X}_{all}|)$ *time complexities, respectively. The intersection of the two sets of features generated by* WILCOXONFILTER *and* ROCFILTER *takes* $O(|V|)$ *time in the worst case [21]. Hence,* WRE *requires* $O(|V|^2|\mathcal{X}_{all}| + i^p \times |V|^3)$ *time. This can be further simplified to* $O(i^p \times |V|^3)$ *since* $|V| > |\mathcal{X}_{all}|$ *in most signaling networks. The time complexity of* WRE-BSE *is* $O(|\mathcal{X}_{all}|^2 \times i^p \times |V|^3)$*, the maximum of the time complexities of* BSE *and* WRE*. The training of the misclassification cost takes* $O((\frac{1}{s} - 1)(i^p \times |V|^3))$*. Hence, taken together,* TRAINMODEL *procedure has time complexity of* $O(O(\mathcal{T}(\cdot)) + (\frac{1}{s} - 1)(i^p \times |V|^3))$*.*

[9] The number of iterations i needed for the grid search is dependent on the number of tuning level l, the range of the parameter to be searched r and the step size during the search s. Formally, it is defined as $i = l \times \frac{r}{s}$.

Network notation	I_1	I_2	I_3	I_4	C_1	C_2	C_3	C_4
Data set (BioModel ID)	MAPK-PI3K (0000000146)	glucose-stimulated insulin secretion (0000000239)	endomesoderm gene regulatory (0000000235)	glucose metabolism (0000000244)	All networks			
Disease node(s)	ERKPP	$\text{ATP}_{mitochondrial}$	Protein_E_Endo16	acetate	{ERKPP, $\text{ATP}_{mitochondrial}$, Protein_E_Endo16, acetate}			
No. of nodes in data set	36	59	622	47	764	764	764	764
No. of hyperedges in data set	34	45	778	109	966	966	966	966
No. (%) of targets in data set	9 (25%)	6 (10.2%)	206 (33.1%)	16 (34%)	237 (31%)	237 (31%)	237 (31%)	237 (31%)
Cross validation	8-fold	5-fold	10-fold	10-fold	10-fold	10-fold	10-fold	10-fold
Test set	Table 6				MAPK-PI3K	glucose-stimulated insulin secretion	endomesoderm gene regulatory	glucose metabolism
No. (%) of targets in test set	1 (25%)	1 (10%)	21 (34.4%)	2 (40%)	9 (25%)	6 (10.2%)	206 (33.1%)	16 (34%)

Table 8: Data set.

Variant	BSE	WRE	WRE-BSE	WMC
TENET-naïve				
TENET-B	√			
TENET-R		√		
TENET-H			√	
TENET-W				√
TENET-WB	√			√
TENET-WR		√		√
TENET-WH			√	√

√ indicates the approach(es) used in the variant.

Weights Ratio ID	C^+	C^-
1	0.1	0.9
2	0.2	0.8
3	0.3	0.7
4	0.4	0.6
5	0.5	0.5
6	0.6	0.4
7	0.7	0.3
8	0.8	0.2
9	0.9	0.1

Table 9: TENET variant and WMC weight ratios used in experiment.

Taken together, the TENET *algorithm requires* $O(|V_{\text{BI}}|^2 + |E_{\text{BI}}| + O(\mathcal{G}(\mathcal{X}_{all})) + O(\mathcal{T}(\cdot) + (\frac{1}{s} - 1)(i^p \times |V|^3)))$ *time for computation. In the worst case, the signaling network is a single strongly connected component with edges connecting every pair of nodes. Such a network implies that* $O(|E_{\text{BI}}|) = O(|V_{\text{BI}}|^2)$*. Hence, in the worst case, the time complexity of* TENET *is* $O((|V| + |E|)^2 + O(\mathcal{G}(\mathcal{X}_{all})) + O(\mathcal{T}(\cdot)))$ *since* $|V_{\text{BI}}| = |V| + |E|$*.*

4 Results and discussion

TENET is implemented using Java. We shall now present the experiments conducted to study the performance of TENET and report some of the results here (additional results are given in Supplementary Material). The experiments are performed on a computer system using a 64-bit operating system with 8GB RAM and a dual core processor running at 3.60GHz. We characterize four signaling networks (referred to as *individual networks*) in *BioModels* (I_1 to I_4 in Table 8) and a *combined network* that is generated by iteratively performing a union of the nodes and edges in individual networks. The resulting combined network is a graph consisting of four disconnected[10] subgraphs, each representing one individual network. For the combined network, we use each of the signaling network as the test set in turn (C_1 to C_4 in Table 8) and examine the effects of generating characterization models from individual networks and from the combined network. Pruning

[10]The node and edge sets of the individual networks are disjoint.

Kernel	Formula	Parameters
Linear	$u^T v$	-
Polynomial	$\gamma(u^T v + c_0)^d$	γ, d, c_0
Radial Basis Function (RBF)	$e^{-\gamma\|u-v\|^2}$	γ
Sigmoid	$\tanh(\gamma u^T v + c_0)$	γ, c_0

Table 10: SVM kernel types and their associated parameters [66].

Parameters	Type	Range Tested
C	SVM parameter	$[2^{-12}$–$2^{12}]$
γ	kernel parameter	$[2^{-12}$–$2^{12}]$
d	kernel parameter	[2–6]
c_0	kernel parameter	$[2^{-12}$–$2^{12}]$

Table 11: SVM kernel types and their associated parameters [66]. Nodes marked with $\sharp$ are known targets.

in TENET is performed on each individual network within the combined network. Section 3.2 describes the generation of the training and test data. We study different variants of TENET (Table 9) by varying the SVM training approach.

4.1 Performance Metrics

We evaluate the performance of TENET based on prediction *accuracy*[11] (ϕ), *sensitivity* (TPR), *specificity* (TNR) and *precision* (PPV) of the generated characterization models using the same training and test set. The definitions are as follows: ϕ=$\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$, TPR=$\frac{\text{TP}}{\text{TP+FN}}$, TNR=$\frac{\text{TN}}{\text{FP+TN}}$ and PPV=$\frac{\text{TP}}{\text{TP+FP}}$ where TP, TN, FP and FN denote true positive, true negative, false positive and false negative prediction, respectively. Note that PPV is set to 0 when the classifier did not make any positive prediction. We include an additional metric *feature reduction factor* (FRF) to compare the performance of the feature selection methods. Formally, FRF=1-$\frac{|\mathcal{F}|}{\mathcal{X}_{all}}$ where $\mathcal{X}_{all}$ is the entire set of features considered in the study. The performance of different characterization models is compared using an *integrated performance score*[12] $\mathcal{P} = \sum_{m \in M} val_m$ where $M = \{\overline{\phi}(val), \phi(test), \text{TPR}, \text{TNR}, \text{PPV}\}$ and val_m is the value of metric m. Note that a larger score indicates better performance.

4.2 Kernel Selection

We experimented with several kernels: linear, radial basis function (RBF), sigmoid and polynomial. The parameters relevant to each kernel type and the ranges of these parameters that we tested are found in Tables 10 and 11, respectively. Figure 6 plots the results of TENET-naïve (which considers all structural features)

[11]The accuracy for the validation and test sets are denoted as $\phi_X(val)$ and $\phi_X(test)$, respectively, where X indicates the method used for training the SVM model. Average prediction accuracy is denoted as $\overline{\phi}$.

[12]This score can be modified according to the needs of the application.

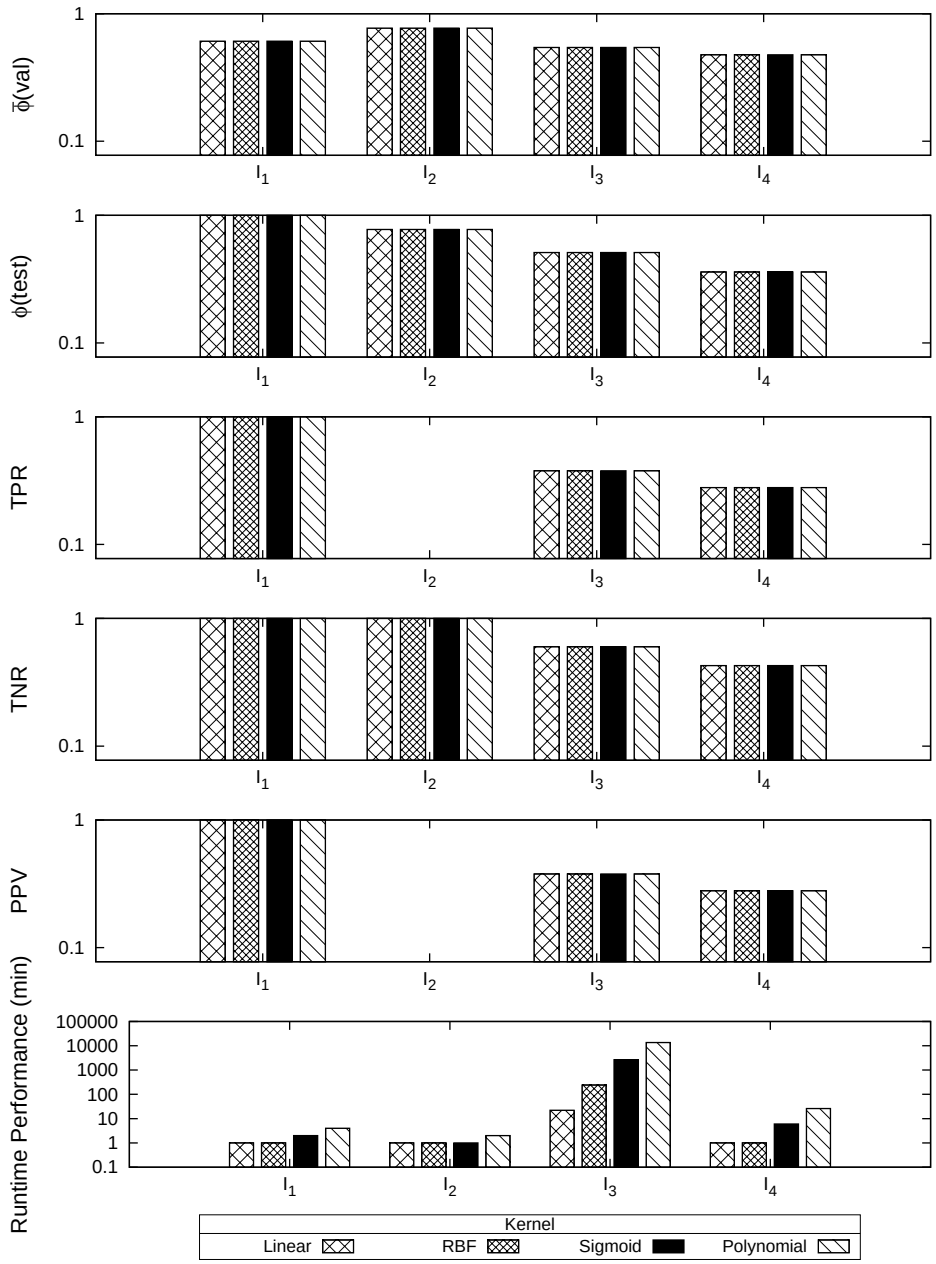

Figure 6: Performance of different kernels using the TENET-naïve approach.

Best model Parameters	I_1	I_2	I_3	I_4
RBF Kernel				
Best model C	$2^{-0.4}$	2^{-10}	$2^{10.8}$	2^{8}
Best model γ	2^{-12}	2^{-10}	2^{-12}	2^{-10}
Sigmoid Kernel				
Best model C	$2^{-0.4}$	2^{-10}	$2^{10.8}$	2^{8}
Best model γ	2^{-12}	2^{-10}	2^{-12}	2^{-10}
Best model $Coef0$	2^{-12}	2^{-10}	2^{-12}	2^{-10}
Polynomial Kernel				
Best model C	$2^{-0.4}$	2^{-10}	$2^{10.8}$	2^{8}
Best model γ	2^{-12}	2^{-10}	2^{-12}	2^{-10}
Best model $Coef0$	2^{-12}	2^{-10}	2^{-12}	2^{-10}
Best model $Degree$	2	2	2	2

Table 12: Best model parameters for the various signaling networks using the TENET-naïve approach with different kernels. The parameters for the TENET-naïve approach with linear kernel is found in Table 13.

using the various kernels. The choice of kernel did not affect the accuracy of the validation and the test sets. This implies that the training data is likely to be linearly separable. The execution time, however, is affected by the number of parameters involved in the kernels and the training set size [80] (size of network). Henceforth, we shall use the linear kernel for the rest of the experiments since it yielded the same accuracy as other kernels but is faster in terms of training speed. The parameters for the best models in this experiment is reported in Table 12. Note that in I_1, the sensitivity (TPR) and precision (PPV) are zero irrespective of the kernels. This highlights a need to use additional techniques (*e.g.*, feature selection) to improve the characterization models. For subsequent experiments, we use the linear kernel as it yielded the same accuracy as other kernels but is faster to train.

4.3 Feature selection

First, we examine the performance of different feature selection approaches (TENET-B, TENET-R and TENET-H) and compare it with TENET-*naïve* for different signaling networks. Note that in this set of experiments, we study the effect of the feature selection approaches in isolation. The effect of incorporating WMC into the SVM shall be investigated later. Table 14 reports the predictive feature sets for each network using different approaches. In total, 24 experiments were conducted since there are three feature selection methods and eight networks (I_1 to I_4 and C_1 to C_4). Amongst these 24 experiments, 25% of the predictive feature sets consist of only one feature while the remaining had multiple features (ranging from 4 to 15 features). This supports our previous observation [16] that *multiple features result in better prediction of known targets*. Observe that in Table 14, bridging centrality is not always in the predictive feature set (*e.g.* , I_2). Figure 7 plots the performances of different feature selection approaches. We can make several observations. First, no single approach performs consistently well on all performance metrics. In fact, network topology plays an important role in feature selection. For instance, I_4 has

Best model C	$\mathbf{I}_1$	$\mathbf{I}_2$	$\mathbf{I}_3$	$\mathbf{I}_4$
TENET-W				
WMC Ratio ID=1	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=2	2^{-10}	2^{-10}	2^{8}	2^{10}
WMC Ratio ID=3	2^{-10}	2^{-10}	$2^{3.2}$	2^{10}
WMC Ratio ID=4	2^{0}	2^{-10}	2^{-2}	2^{10}
WMC Ratio ID=5 (TENET-naïve)	$2^{-0.4}$	2^{-10}	$2^{10.8}$	2^{8}
WMC Ratio ID=6	$2^{-4.08}$	2^{-10}	$2^{2.8}$	$2^{5.6}$
WMC Ratio ID=7	$2^{-5.76}$	2^{-10}	$2^{0.08}$	2^{6}
WMC Ratio ID=8	2^{0}	2^{-10}	$2^{5.6}$	$2^{6.8}$
WMC Ratio ID=9	$2^{0.8}$	2^{6}	2^{8}	2^{8}
TENET-WB				
WMC Ratio ID=1	$2^{7.28}$	2^{-10}	2^{-10}	$2^{8.4}$
WMC Ratio ID=2	$2^{6.4}$	2^{-10}	2^{8}	$2^{11.6}$
WMC Ratio ID=3	2^{6}	2^{-10}	$2^{0.16}$	$2^{10.8}$
WMC Ratio ID=4	$2^{2.4}$	2^{-10}	2^{-2}	$2^{10.4}$
WMC Ratio ID=5 (TENET-B)	2^{4}	2^{-10}	$2^{8.8}$	2^{12}
WMC Ratio ID=6	2^{4}	2^{-10}	2^{6}	$2^{10.4}$
WMC Ratio ID=7	2^{2}	2^{-10}	$2^{0.8}$	$2^{8.4}$
WMC Ratio ID=8	2^{4}	2^{-10}	2^{10}	$2^{11.2}$
WMC Ratio ID=9	$2^{4.8}$	2^{4}	$2^{7.6}$	2^{10}
TENET-WR				
WMC Ratio ID=1	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=2	2^{-10}	2^{-10}	2^{6}	2^{-10}
WMC Ratio ID=3	2^{-10}	2^{-10}	$2^{7.52}$	2^{-10}
WMC Ratio ID=4	2^{4}	2^{-10}	$2^{3.6}$	2^{-10}
WMC Ratio ID=5 (TENET-R)	2^{6}	2^{-10}	$2^{8.4}$	2^{-10}
WMC Ratio ID=6	$2^{-3.6}$	2^{-10}	$2^{-5.52}$	2^{-10}
WMC Ratio ID=7	$2^{-3.2}$	2^{0}	$2^{-0.24}$	2^{4}
WMC Ratio ID=8	21.84	2^{-10}	$2^{4.96}$	2^{-10}
WMC Ratio ID=9	2^{10}	2^{8}	2^{6}	2^{-10}
TENET-WH				
WMC Ratio ID=1	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=2	2^{6}	2^{-10}	2^{6}	2^{-10}
WMC Ratio ID=3	2^{4}	2^{-10}	$2^{6.8}$	2^{-10}
WMC Ratio ID=4	$2^{8.08}$	2^{-10}	$2^{2.4}$	2^{-10}
WMC Ratio ID=5 (TENET-H)	2^{10}	2^{8}	$2^{11.6}$	2^{-10}
WMC Ratio ID=6	2^{10}	2^{2}	2^{4}	2^{-10}
WMC Ratio ID=7	2^{10}	2^{2}	$2^{-1.76}$	2^{4}
WMC Ratio ID=8	2^{10}	2^{4}	$2^{6.4}$	2^{-10}
WMC Ratio ID=9	2^{8}	2^{10}	2^{4}	2^{-10}

Table 13: Best model C parameter for the various signaling networks using various approaches with linear kernel.

Data	TENET-B	TENET-R	TENET-H
I_1	δ,π,θ_{in},θ_{out}	δ,ζ,β,ϑ,θ_{out},μ,κ_{undir}	δ,ζ,β,ϑ
I_2	θ_{in}	δ,π,β,κ_{undir},κ_{cyc},α,θ_{in},κ_{in},μ,κ_{mid},θ_{out},θ_{total}	π,β,κ_{cyc},κ_{undir}
I_3	δ,ζ,π,β,κ_{cyc},ϑ,α,κ_{in},κ_{mid},μ,κ_{out},θ_{out},ω,θ_{total},κ_{undir}	δ,ζ,ϑ,α,κ_{mid},θ_{out}, θ_{total},ω,κ_{undir}	δ,ζ,ϑ,α,θ_{out}, θ_{total},κ_{undir}
I_4	ζ,β,κ_{cyc},ϑ,α,κ_{in},κ_{mid},μ,ω,κ_{out},θ_{out},θ_{total}, κ_{undir}	ω	ω
C_1	δ,ζ,π,β,κ_{cyc},ϑ,α,θ_{in}, κ_{mid},θ_{out},μ,ω,κ_{undir}	δ,ζ,π,β,ϑ,α,κ_{mid},κ_{undir}, θ_{out}	ζ,π,ϑ,α,θ_{out}, κ_{undir}
C_2	δ,ζ,π,β,κ_{cyc},ϑ,α,θ_{in}, κ_{mid},θ_{out},κ_{undir}	δ,ζ,α,κ_{mid},θ_{out},ω,θ_{total}, κ_{undir}	δ,ζ,α,κ_{mid},ω, κ_{undir}
C_3	θ_{in}	ζ	ζ
C_4	ζ,π,β,κ_{cyc},ϑ,α,κ_{in},θ_{in},ω,κ_{out},θ_{out},θ_{total}, κ_{undir}	δ,ζ,π,ϑ,α,κ_{mid},θ_{out},ω, θ_{total},κ_{undir}	ζ,π,ϑ,α,κ_{undir},ω,θ_{out},θ_{total},κ_{mid}

Table 14: Features selected by various feature selection approaches.

extremely high density of edges (ratio of edges to nodes) compared to other net-

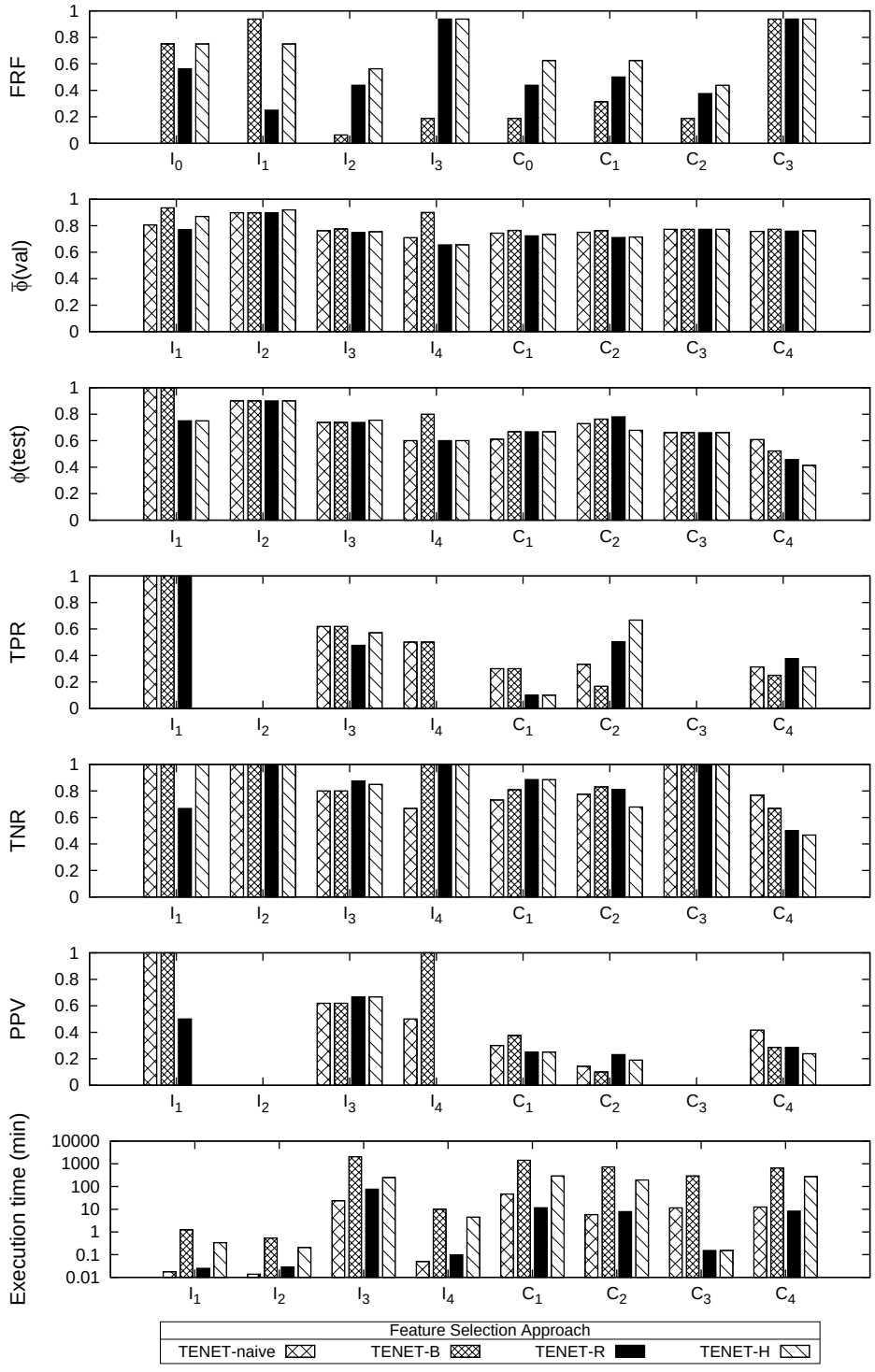

Figure 7: Performance of different feature selection approaches.

works. The connectivity features of such networks become less informative and other features such as target downstream effect becomes more important. Hence, *the most appropriate feature selection approach is dependent on the signaling network*. However, we note that for larger sized networks, a larger number of features are informative (regardless of feature selection approach). This is perhaps because larger networks provide greater richness of context and diversity of structure in the sub-networks. Since network sizes are growing and network analysis demands applicability to larger networks, future methods might benefit particularly from the use of multiple features. Second, *feature selection generally led to an improvement in prediction accuracy* (87.5% for validation data set and 50% in test data set) over the naïve approach. An exception is C_4 in which feature selection resulted in poorer performance. In C_4, the characterization model is generated using I_1, I_2 and I_3 as training data whereas I_4 is used as the test data. The characteristics of the known targets in the training data may be quite different from that of the test data. Indeed, from Table 14, we observe that bridging coefficient π is included in the predictive topological feature set of C_4, but not in I_4. Including redundant features may lead to poorer performance. Third, the models generally have high specificity due to imbalanced data set. Fourth, TENET-R has the best runtime performance, followed by TENET-H and TENET-B. The poorer performance of TENET-B is due to the interaction of the feature selection approach with the classifier (classifier-aware approach) which is different from TENET-R where the feature selection approach is a wrapper layer that sits on top of the classifier. Finally, the size of the networks used for training affects the runtime performance. In general, larger size networks require longer runtime. In Section 4.7, we report TENET's performance on the human cancer signaling network containing more than 2500 nodes.

4.4 Effect of varying WMC

Intuitively, when we vary the WMC, we expect that as the target misclassification cost C^+ increases, the prediction accuracy, sensitivity, specificity and precision would display a negative skewed, increasing, decreasing and positive skewed distribution, respectively. This is because a large C^+ eventually results in a model that is likely biased towards classifying data as targets. The effect of varying the WMC are reflected in Figures 8 to 15. From the figures, we noted the following trends. First, amongst the individual networks, only I_3 (Figure 10) displays the expected trends. This could be due to the extreme small target size (1 or 2) in the test set that resulted in extreme fluctuations in the performance metrics and deviation from the expected trends. Hence, the target size of the test set can have significant impact on the observed results. Second, the performance of the combined networks C_1 (Figure 12), C_2 (Figure 13) and C_4 (Figure 15) resembles that of I_3, possibly due to the large size of I_3 dominating over other networks used for training. This implies large training networks can have undue influence on the characterization model. Third, sensitivity generally improves whereas specificity generally deteriorates when the target misclassification cost is set higher than the non-target

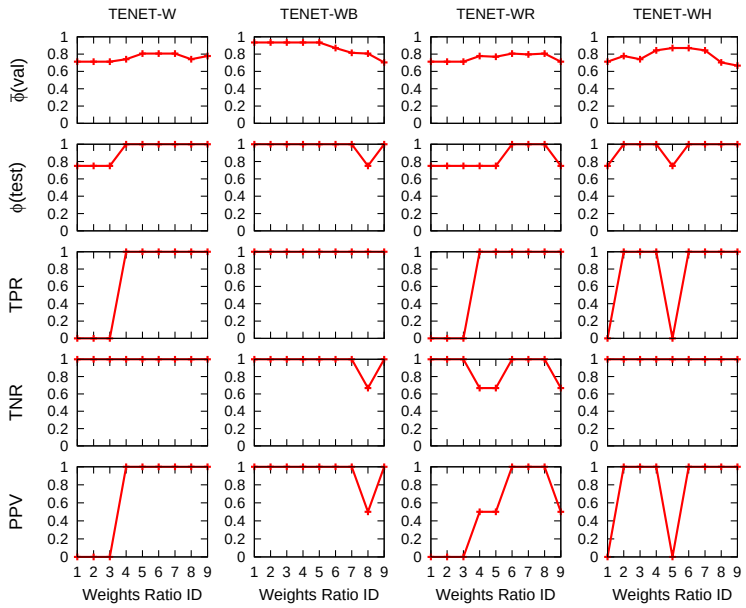

Figure 8: Performance of TENET variants incorporating feature selection approach and WMC for the MAPK-PI3K network.

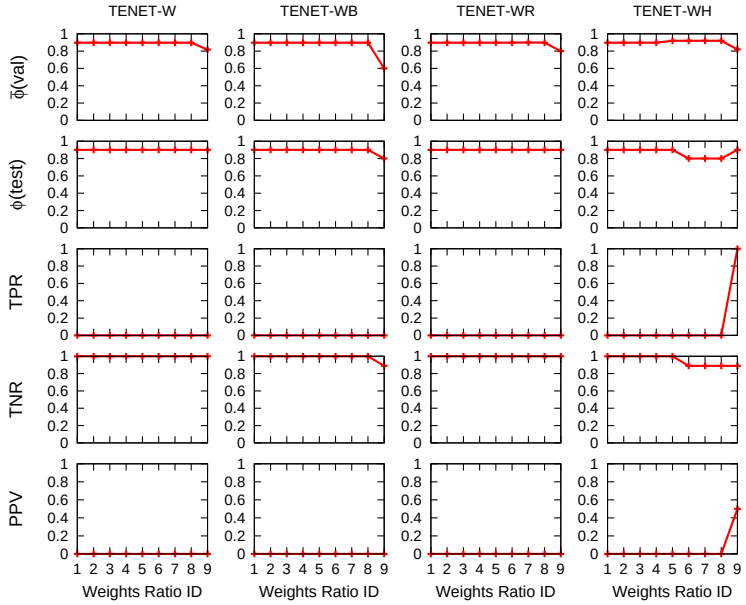

Figure 9: Performance of TENET variants incorporating feature selection approach and WMC for the glucose-stimulated insulin secretion network.

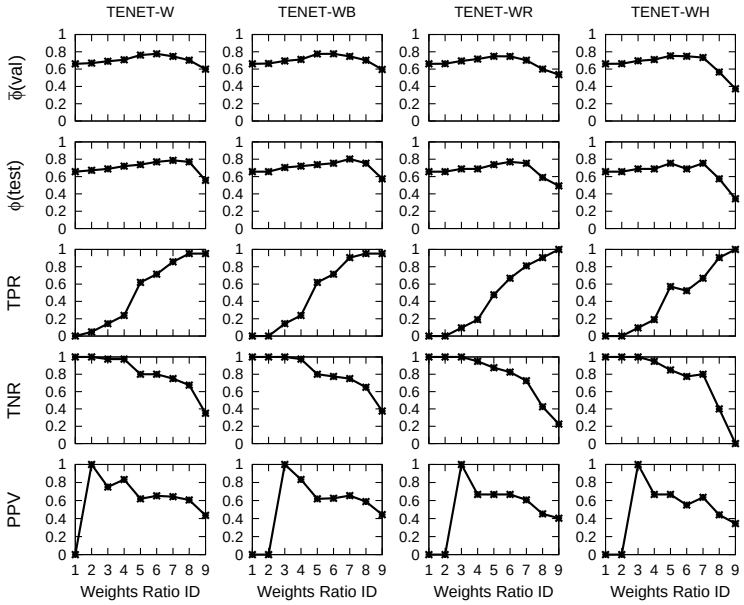

Figure 10: Performance of TENET variants incorporating feature selection approach and WMC for the `endomesoderm` gene regulatory network.

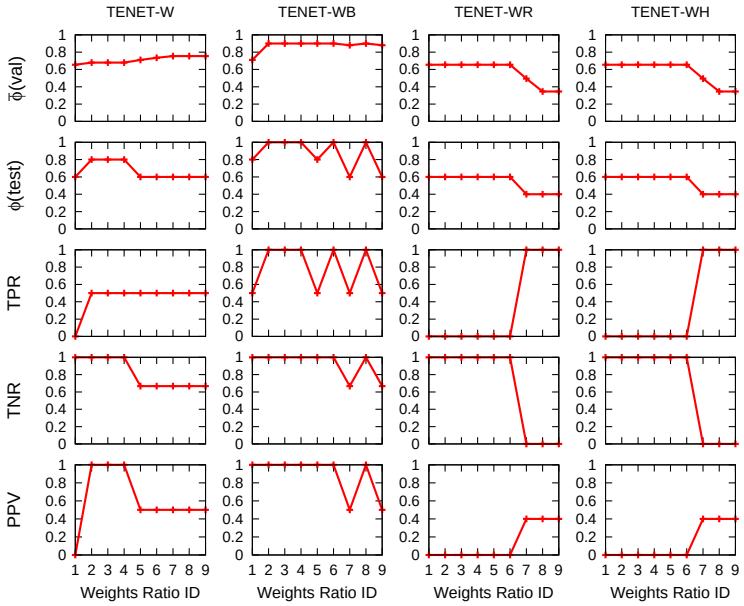

Figure 11: Performance of TENET variants incorporating feature selection approach and WMC for the `glucose` metabolism network.

misclassification cost ($C^+ > C^-$). The choice of an appropriate model depends on the application. Fourth, the prediction accuracy tends to display a skewed distribution where accuracy initially increases (or remains constant) with increasing C^+,

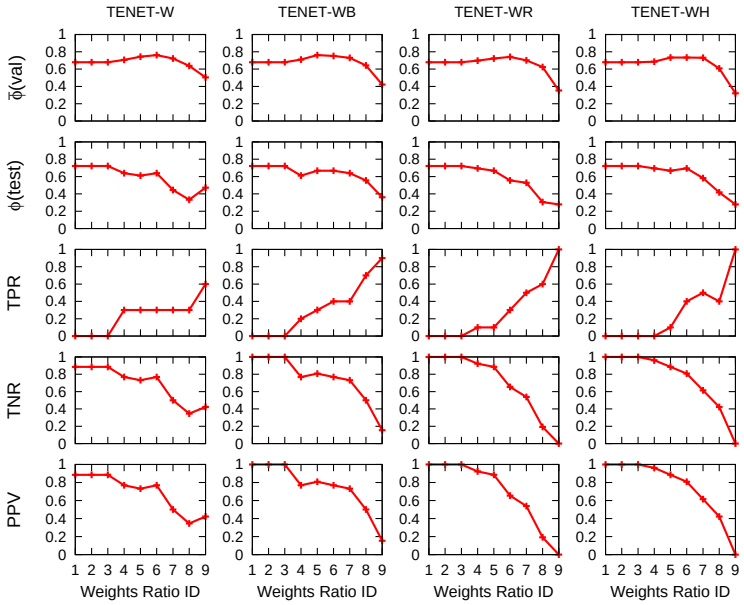

Figure 12: Performance of TENET variants incorporating feature selection approach and WMC for C_1.

and then decreases with increasing C^+. Fifth, individual networks and combined networks behave differently. In individual networks, prediction accuracy, sensitivity and precision generally improve when C^+ is set larger than C^-. However, in combined networks, sensitivity improves whereas other performance criteria deteriorates when C^+ is set larger than C^-. Hence, *there is no single universal best value of C^+ and the choice of C^+ depends on the network.*

4.5 Best TENET variant

We identify the best TENET variant (Table 15) using the integrated performance score $\mathcal{P}$. We note the following. First, the best TENET variant is network dependent. Second, *variants incorporating both* WMC *and feature selection generally perform well*. Specifically, setting C^+ greater than C^- led to better results. Third, TENET *variants based on individual networks (I_1 to I_4) outperforms that based on combined networks* (C_1 to C_4). The poorer performance of the combined networks may be due to insufficient number of training networks, inappropriate or insufficient features used for training or that signaling networks by nature have distinct characteristics and it is just not possible to have a generalized model. Finally, the predictive topological features differ across networks (Tables 14 and 15). Hence, as we mentioned in Section 1, *a single set of predictive topological features may not effectively characterize known targets in all signaling networks*. When we compare the results with that in our previous work, we note that the set of predictive topological features are different from the discriminative topological features

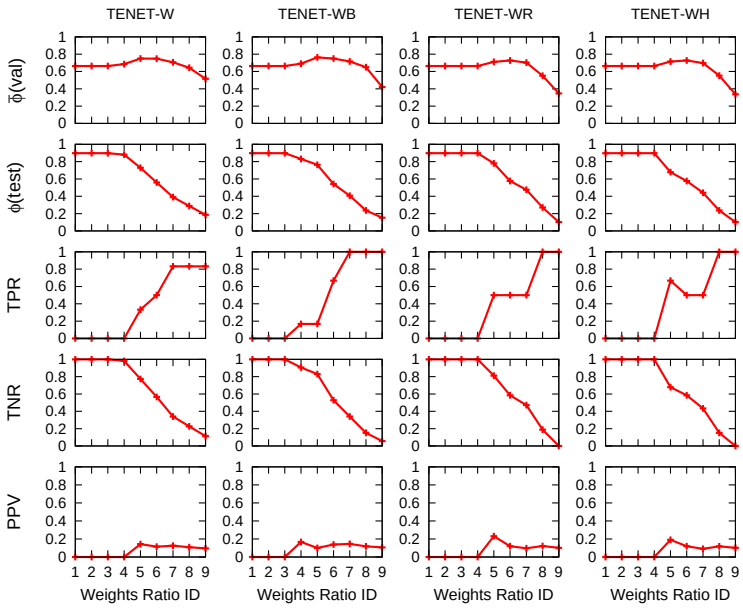

Figure 13: Performance of TENET variants incorporating feature selection approach and WMC for C_2.

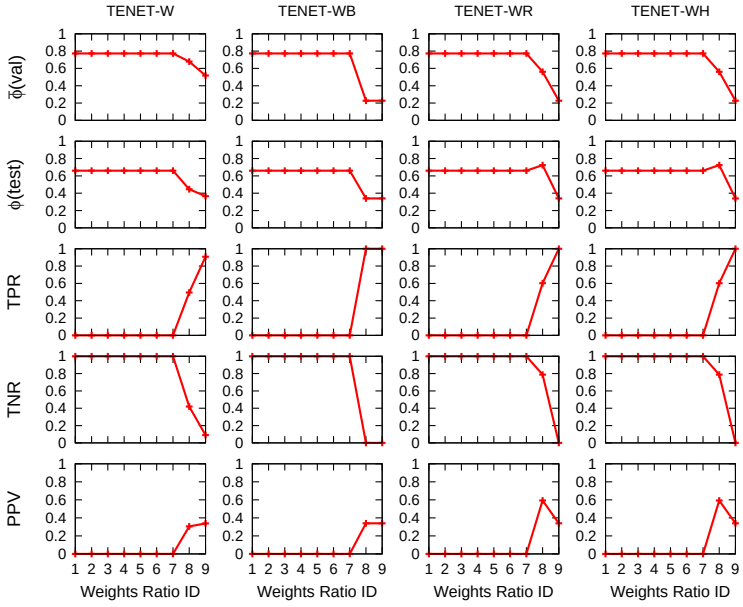

Figure 14: Performance of TENET variants incorporating feature selection approach and WMC for C_3.

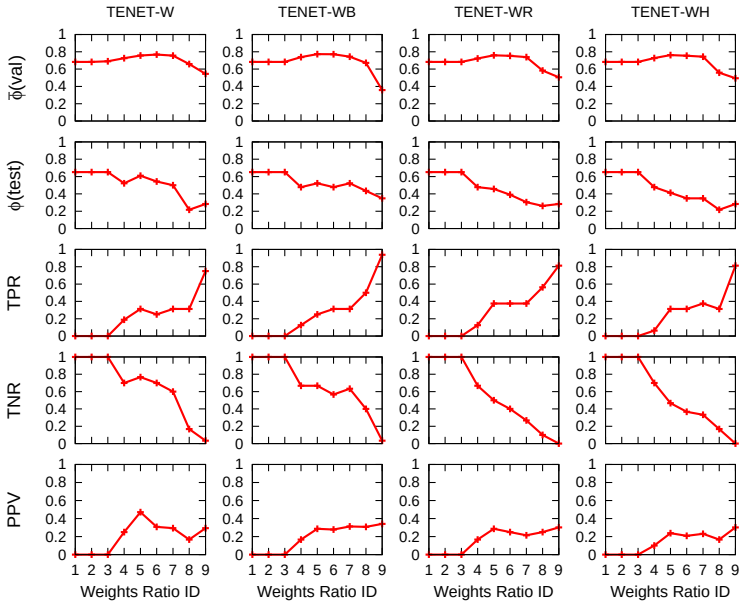

Figure 15: Performance of TENET variants incorporating feature selection approach and WMC for C_4.

	I_1	I_2	I_3	I_4	C_1	C_2	C_3	C_4
Best Approaches	TENET-B$^\sharp$, TENET-WB (C^+=0.1,0.2,0.3, 0.4)	TENET-WH (C^+=0.9$^\sharp$)	TENET-WB (C^+=0.7$^\sharp$)	TENET-WB (C^+=0.2,0.3,0.4, 0.6,0.8$^\sharp$)	TENET-WH (C^+=0.6$^\sharp$)	TENET-R$^\sharp$	TENET-WR (C^+=0.8$^\sharp$), TENET-WH (C^+=0.8)	TENET-naïve$^\sharp$
$\mathcal{P}$	4.935	4.109	3.86	4.9	3.08	3.022	3.268	2.917
$\bar{\phi}(val)$ **[**$\Delta\bar{\phi}(val)$**]**	0.935 [0.16]	0.82 [-0.087]	0.747 [-0.02]	0.9 [0.268]	0.734 [-0.013]	0.711 [-0.052]	0.561 [-0.274]	0.757 [0]
$\phi(test)$ **[**$\Delta\phi(test)$**]**	1 [0]	0.9 [0]	0.803 [0.088]	1 [0.667]	0.694 [0.136]	0.78 [0.070]	0.724 [0.097]	0.609 [0]
TPR [Δ**TPR]**	1 [0]	1 [∞^*]	0.905 [0.462]	1 [1]	0.4 [0.333]	0.5 [0.502]	0.602 [∞^*]	0.313 [0]
TNR [Δ**TNR]**	1 [0]	0.889 [-0.111]	0.75 [-0.063]	1 [0.499]	0.808 [0.105]	0.811 [0.048]	0.788 [-0.212]	0.767 [0]
PPV [Δ**PPV]**	1 [0]	0.5 [∞^*]	0.655 [0.058]	1 [1]	0.444 [0.48]	0.231 [0.615]	0.593 [∞^*]	0.471 [0]

Table 15: Summary of best TENET variant for different networks. C^+ values are provided in bracket besides approaches using WMC. $\Delta_x = \frac{x_{best} - x_{naïve}}{x_{naïve}}$ where x_{best} and $x_{naïve}$ are the values of performance metric x of the best TENET variant and TENET-naïve, respectively. * marks instances where $x_{naïve} = 0$ and $^\sharp$ marks the best models selected for generating the characterization model.

(DTF) identified in [16] although there was an overlap of at least 50% of the features[13]. The difference is due to the different approach used to identify the features. The characterization models[14] generated by these DTFs also yielded poorer average ROC (0.873) than that generated using TENET (0.913) (Approach DIFFER in Figure 16).

[13]We consider only I_1 to I_3 and exclude I_4 from this comparison as no DTF was found at p-value less than 0.05

[14]We use SVM with WMC and WRE to generate the characterization models.

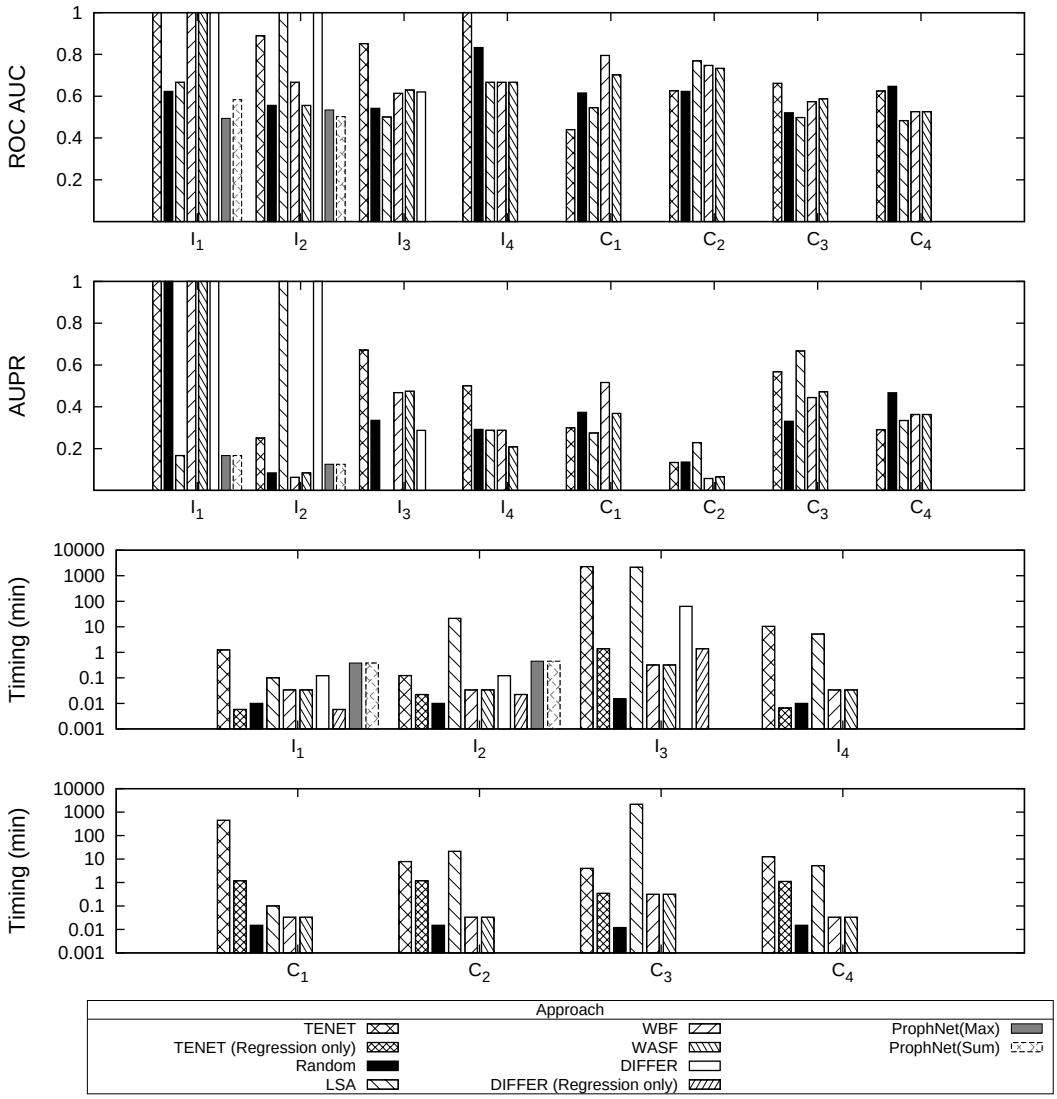

Figure 16: Performance of different prioritization approaches.

4.6 Comparison with state-of-the-art approaches

In this subsection, we describe the experiments that compare TENET against other state-of-the-art approaches. We perform three sets of experiments for comparison with *network-unaware* techniques, PPI network-based techniques and *network-aware* target prioritization approaches. Recall that state-of-the-art techniques such as [41, 65, 108] focus on PPI networks instead of signaling networks. To the best of our knowledge, there does not exist any target characterization technique for signaling networks. However, one way to investigate the performance of TENET is to examine how well the characterization model generated by it *prioritizes* known targets. Intuitively, *target prioritization* aims to *rank* the nodes according to their

Network	Targets in review article	Total targets	% of TENET top-10 ranked nodes overlapping with review targets
I_1	RAB25, PRKCI, EVI1, PIK3CA, FGF1, MYC, PIK3R1, AKT2♯, AURKA, KRAS♯, BRAF♯, CTNNB1, CDKN2A, APC, KIT, SMAD4, IGF2, SAT2, ARHI, PEG3, PLAGL1, RPS6KA2, TP53, BRCA1, BRCA2, PTEN, OPCML, WWOX, DAPK1, CDH13, MLH1, ICAM1, DNAJC15, MUC2, PCSK6, CDKN1A, RASSF1, SOCS1, SOCS2, PYCARD, SFN	41	50%
I_2	GLP-1, GLP-1 receptor, DPP-4, NEP-24.11, SGLT, amylin, PPAR, ATP-sensitive potassium channel, α-glucosidase, glucokinase♯, AMP kinase♯, carnitine palmitoyltransferase-1, glycogen synthase kinase-3, PTP-1B, pyruvate dehydrogenase♯, fructose-1,6-bisphosphatase♯, 11β-hydroxysteroid dehydrogenase 1, sirtuin 1, acyl-CoA-diacylglycerol acyltransferase 1, phosphoenolpyruvate carboxykinase♯, glucose-6-phosphatase, PPARγ coactivator 1α, acetyl-CoA carboxylase♯, mitochondrial rotenone-sensitive NADH:ubiquitone oxidoreductase (complex I)♯, leptin, ghrelin, resistin, C-peptide, protein kinase C, AGE, RAGE, glutamine:fructose-6-phosphate♯, PARP, VEGF, aldose reductase♯, vitamin C, vitamin E, GPR40, GPR119, GPR41, GPR43, GPR120, GPR109A, dopamine-2 receptor, m3 subtype muscarinic receptor, 5-hydroxytryptamine 2c subtype serotonin receptor, imidazoline, glucagon receptor, retinoid X receptor, colesevalam, IL-1β, chemokine receptor 2, angiotensin receptor, thioredoxin-interacting protein, Kv2.1 channel, FBF21, $\omega-3$ PUFA, ZnT8, diacylglycerol acyl transferase 1	60	50%
I_3	GSK-3♯, frizzled♯, nβ-TCF♯, HesC♯, Wnt8♯, Hox11/13b♯, Su(H)♯, Blimp1♯, Otx♯, Bra♯, FoxA♯, GataE♯, Gcm♯, Notch♯	14	90%
I_4	ack, pta, acs♯, poxB♯, pykA, pykF♯, fadR, ppc♯, pyc, zwf, PTS♯, galP, glucokinase♯, glucose♯	14	50%

Table 16: Summary of targets obtained from review articles. Targets that are present in the network and in TENET top-10 ranked nodes are marked as ♯ and underlined, respectively.

potential of being a target based on some *importance measures* (*e.g.*, gene expression level [14]. In the following, we first describe the comparisons with *network-unaware* techniques, then that with PPI networks and finally that with the *network-aware* target prioritization approaches.

4.6.1 Comparison with network-unaware approaches

We compared TENET's prediction against those derived from non-network-based approaches, specifically, targets that were predicted by various experimental techniques and consolidated within review articles. The targets for I_1, I_2, I_3 and I_4 were derived from [3], [97], [60] and [87], respectively. Table 16 summarizes the targets found in the review articles (referred to as *review targets*). In general, there is an overlap between the targets in the network and those in the reviews. Note that in signaling networks, the same gene and protein often exist in multiple forms and such representations may be manifested in the top-10 ranked nodes. For instance,

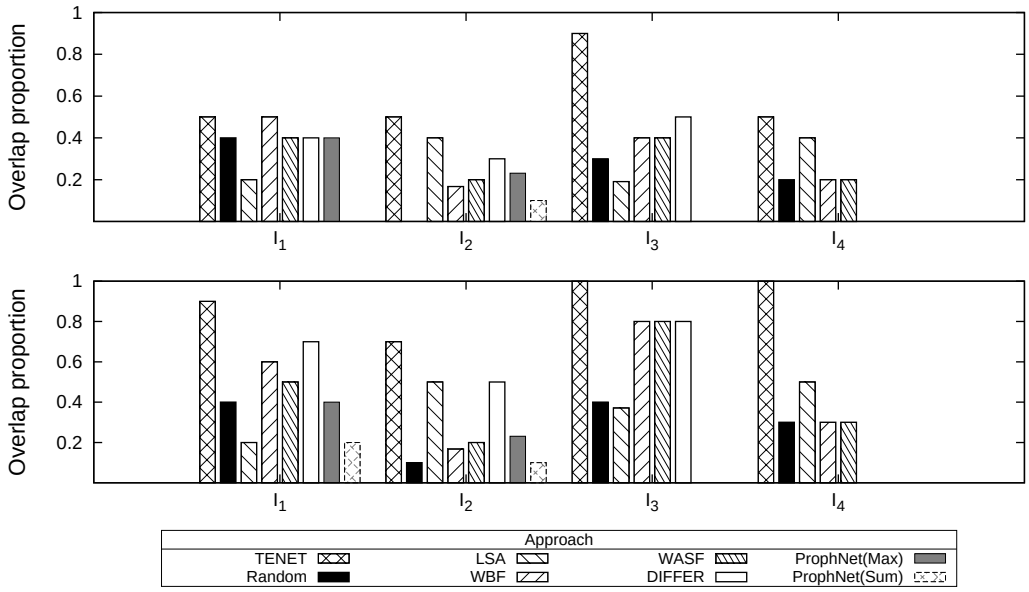

Figure 17: Proportion of top-10 ranked nodes that overlaps with review targets (top) and review and curated targets (bottom)

in TENET top-10 ranks of I_3, there are multiple versions of nβ-TCF[15]. When we take this multiple forms into consideration, the percentage of TENET top-10 ranks overlapping with review targets are 50%, 50%, 90% and 50% for I_1, I_2, I_3 and I_4, respectively. Next, we examine if the remaining targets are biologically relevant by checking for correspondence with our curated targets in Section 2.3. Only three targets[16] do not correlate with curated targets. This implies that top ranking nodes in TENET correlate well with existing biological knowledge. That is, there is good correspondence between TENET top-10 ranked nodes and existing biological knowledge. Note that TENET also out-performs other state-of-the-art approaches in terms of the overlap achieved between the top-10 ranked nodes and the review targets and curated targets (Figure 17).

4.6.2 Comparison with PPI-based approaches

Following which, we compared TENET with several PPI-based target prioritization tools, namely, *NetworkPrioritizer* [47], *ToppGene* [12] and *ProphNet* [64]. The comparison with *NetworkPrioritizer* is presented and discussed in the main text. The *ToppNet* feature in *ToppGene* is used to prioritize the nodes.

Note that *ToppNet* requires a set of training and test nodes as inputs for analysis. These node sets have to be provided as either sets of HGNC, *Entrez*, *Ensembl*,

[15] nβ-TCF is present as protein P nβ-TCF, protein M nβ-TCF and protein E nβ-TCF which represent nβ-TCF in protein form in PMC, mesoderm and endoderm cells, respectively.

[16] PP2A in I_1 and ferricytochrome c, dihydroxyacetone-phosphate and succinyl-CoA from I_2

RefSeq or *UniProt* identifiers. We annotate the nodes in the four networks using *UniProt* and *Entrez* identifiers when majority of the nodes are proteins and genes, respectively. For clarity, networks I_1, I_2 and I_4 are annotated using *UniProt* whereas I_3 is annotated using *Entrez*. We follow the following rules during annotations:

1. When multiple annotation are available, select the one with organism matching that of the given network.

2. When a node does not have a valid annotation, it inherits an annotation (where available) related to an edge that the node is associated with. For example, `glucose` do not have a corresponding *UniProt* identifier. It is involved in a reaction where `glucokinase` (*UniProt* ID=P52792) catalyzes `glucose` to `glucose-6-phosphate`. Hence, it shall inherit the *UniProt* identifier of `glucokinase`.

3. When a node is a result of post-translational modification (*e.g.*, phosphorylation), it inherits the annotations of the original node. For example, phosphorylated `ERK` shall have the same annotations as unphosphorylated `ERK`.

ToppNet, which prioritizes nodes based on functional annotations, returns no results for all four networks and is excluded from Figure 16. This could imply that the database[17] in *ToppNet* are lacking in functional annotations related to these networks. Unlike *ToppNet* whose analysis is reliant on the quality of its functional annotations database, TENET analysis depends only on the structure of the network which is inherent in the signaling network given by the user.

ProphNet provides several prioritization features. In particular, we are interested in prioritization of a given set of nodes and prioritization of a *ProphNet*-generated node set for a given disease and we refer to them as *ProphNet* A and *ProphNet* B, respectively. *ProphNet* A returns no results for I_1 to I_4 whereas *ProphNet* B returns a prioritized *ProphNet*-generated node set for I_1 and I_2 when "ovarian fibromata" and "diabetes mellitus, insulin-dependent, 2" were given as the input disease, respectively. Note that for I_3 and I_4, we were not able to find a related disease tag in *ProphNet* B for a meaningful query. The *ProphNet*-generated and prioritized node sets (referred to as *ProphNet* nodes) where mapped to the nodes in I_1 and I_2. The mapping is performed according to the following rules:

1. When a node does not have a clear, unambiguous 1-1 *ProphNet* node mapping, it is mapped to a related *ProphNet* node. For example, `GS` (`Grb2-sos`) is mapped onto *ProphNet* nodes `GRB2`, `sos1` and `sos2` and inherit the *ProphNet* score of these nodes.

2. When multiple *ProphNet* node versions are available, all *ProphNet* nodes are mapped to the network node. For example, different *ProphNet* nodes

[17] *ToppNet* database contain human and mouse genes and uses *GO* as functional annotations whereas I_1 to I_4 are networks related to mouse (I_1 and I_2), sea urchin (I_3) and `E. Coli` (I_4).

such as PPP2R5E, PPP2R3A and PPP2R2B are used to represent different components of PP2A. All of these *ProphNet* nodes and their values are mapped to PP2A.

This mapping resulted in some nodes being mapped to one or more *ProphNet* nodes. Ambiguity in the node prioritization occurs when a node is mapped to multiple *ProphNet* nodes. We resolve this ambiguity by generating two new set of prioritization ranks called *ProphNet(Max)* and *ProphNet(Sum)*. In the former rank, a node will be assigned the highest *ProphNet* ranks among the mapped *ProphNet* nodes. In the latter rank, a node will be given a score that is the sum of the *ProphNet* values for all the mapped *ProphNet* nodes. Then, the nodes are ranked in decreasing order of this score. Note that node mappings have to be performed either during annotations for *ToppNet* or when comparing the prioritized *ProphNet* nodes. Ambiguity arise when the signaling networks contain the same protein or gene in different forms or cells as different nodes whereas the tools expect each node to represent a different protein or gene. Hence, a fair comparison between TENET and these two approaches becomes difficult. Note that in both I_1 and I_2 (Figure 16), TENET outperforms *ProphNet* in terms of ROC AUC and AUPR. For execution time, TENET also outperforms *ProphNet* when training was performed offline. Hence, in subsequent experiments, we shall focus on comparing TENET and other signaling network-based approaches (random prioritization, DIFFER, LSA and *NetworkPrioritizer*).

4.6.3 Comparison with network-aware target prioritization approaches

Target prioritization is the process of ranking nodes in a network according to their likelihood of being a target based on some criteria (*e.g.*, sensitivity, gene expression level, score generated by a characterization model). That is, given a signaling network $G = (V, E)$, the **target prioritization problem** assigns a **target rank** r_u for each node $u \in V$. Given $u, v \in V$, u is more likely than v to be a target if $r_u < r_v$. It is potentially useful in helping to plan experiments since resources are limited and experiments can be costly and time-intensive. This is especially true in drug development [68]. Note that target characterization do not generate "new" targets, but instead produces a model that characterizes known targets. In contrast, target prioritization may generate "new" targets by virtue of the fact that high ranking nodes that are not in the set of known targets have greater potential to be "new" targets. Characterization models (described in Section 3.1) generated by TENET can be used for generating *prioritization scores* to rank nodes in a signaling network. Recall that TENET generates characterization models using an approach based on support vector machines (SVM). SVM (*e.g.*, ϵ-support vector regression) can yield models that produce a continuous outcome (*i.e.*, regression scores) instead of discrete outcome (*i.e.*, classes). Hence, the regression scores can then be used for prioritizing nodes. Figure 18 depicts the overview of the target prioritization process using the output of TENET. Specifically, TENET produces a characterization model when given a signaling network (*e.g.*, MAPK/PI3K network),

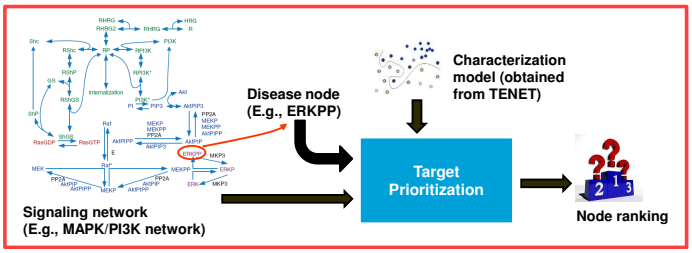

Figure 18: Target prioritization using TENET.

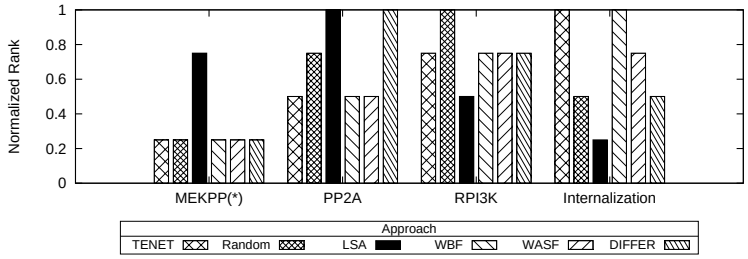

Figure 19: Normalized rank of nodes in test set of I_1.

a relevant disease node (*e.g.*, `ERKPP`) and a list of known targets (*e.g.*, `Akt`, `Raf`, *etc.*) as inputs. This characterization model can then be applied on any signaling network G to obtain a regression score for each node in G. Finally, the nodes in G are ranked in decreasing order of the regression score. Nodes that are ranked top but not in the set of known curated targets may be potential "new" targets.

For our study, we compare TENET with several *network-aware* target prioritization approaches, namely, random prioritization, LSA [33] and *NetworkPrioritizer* [47]. In random prioritization, the nodes were randomly assigned a rank in the range [1–$|V|$] where $|V|$ is the number of nodes in the network and we assume that no ranking ties are present. LSA was performed using *Copasi* [85] with the following configuration: {task=sensitivities; subtask=time series; function=all variables of the model; and variable=all parameter values}. We consider both *Weighted Borda Fuse* (WBF) and *Weighted AddScore Fuse* (WASF) in *NetworkPrioritizer* and consider all features provided. Note that uniform weights were used for rank aggregation since we do not have prior knowledge of the best weights or features to consider. For TENET, we use the characterization model to generate prioritization ranks of known targets. Specifically, we apply the SVM models to obtain these ranks. The SVM type is set to ϵ-support vector regression (ϵ-SVR)[18] with default ϵ value (1x10^{-3}) and the SVM parameters are set according to the best models for each network (Tables 13, 15 and 17). Note that the nodes are ranked in decreasing order of the regression score and higher ranked nodes are more likely to be targets.

First, the experimental results reveal that the *normalized ranks* of a given node vary widely using different approaches. Figures 19, 20 and 21 plot the *normalized*

[18] In ϵ-SVR, the error function is an ϵ-insensitive loss function and error smaller than ϵ is ignored [8].

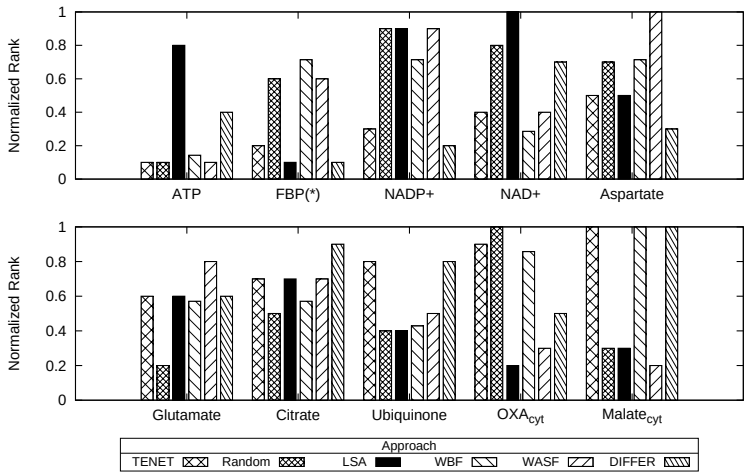

Figure 20: Normalized rank of nodes in test set of I_2.

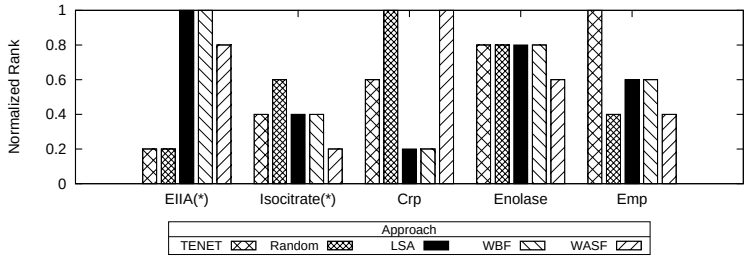

Figure 21: Normalized rank of nodes in test set of I_4.

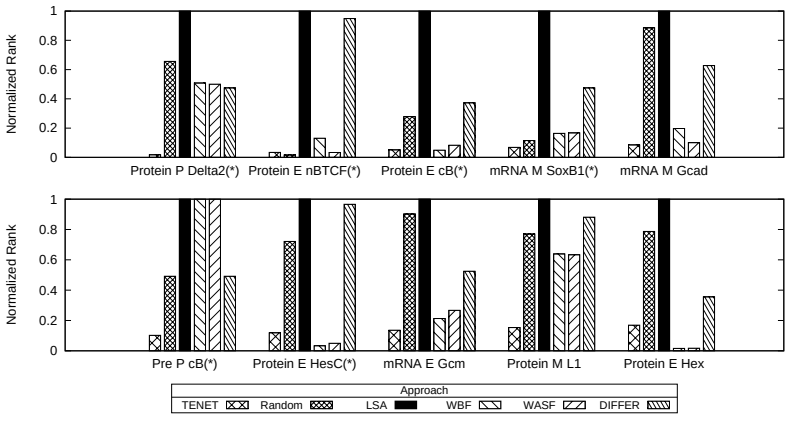

Figure 22: Normalized rank of top 10 nodes ranked using TENET in test set of I_3.

Best model C	C_1	C_2	C_3	C_4
TENET-W				
WMC Ratio ID=1	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=2	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=3	2^{-10}	2^{-10}	2^{-10}	2^{2}
WMC Ratio ID=4	$2^{0.4}$	$2^{-1.92}$	2^{-10}	$2^{9.2}$
WMC Ratio ID=5 (TENET-naïve)	$2^{10.4}$	$2^{5.6}$	2^{-10}	$2^{7.6}$
WMC Ratio ID=6	2^{0}	$2^{11.92}$	2^{-10}	$2^{1.04}$
WMC Ratio ID=7	$2^{-0.64}$	$2^{10.4}$	2^{-10}	$2^{-0.08}$
WMC Ratio ID=8	2^{8}	$2^{6.4}$	$2^{2.8}$	2^{8}
WMC Ratio ID=9	$2^{9.6}$	2^{6}	2^{6}	2^{8}
TENET-WB				
WMC Ratio ID=1	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=2	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=3	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=4	2^{4}	$2^{8.8}$	2^{-10}	$2^{3.6}$
WMC Ratio ID=5 (TENET-B)	$2^{9.2}$	$2^{5.36}$	2^{-10}	$2^{7.6}$
WMC Ratio ID=6	$2^{6.4}$	2^{4}	2^{-10}	$2^{-0.32}$
WMC Ratio ID=7	2^{2}	$2^{6.4}$	2^{-10}	$2^{10.8}$
WMC Ratio ID=8	$2^{4.4}$	$2^{9.6}$	2^{-10}	2^{8}
WMC Ratio ID=9	2^{6}	2^{6}	2^{-10}	$2^{7.6}$
TENET-WR				
WMC Ratio ID=1	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=2	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=3	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=4	2^{4}	2^{-10}	2^{-10}	2^{8}
WMC Ratio ID=5 (TENET-W)	2^{0}	2^{6}	2^{-10}	$2^{7.6}$
WMC Ratio ID=6	$2^{-0.16}$	$2^{-5.76}$	2^{-10}	2^{-2}
WMC Ratio ID=7	$2^{0.70}$	$2^{-3.6}$	2^{-10}	2^{10}
WMC Ratio ID=8	$2^{0.62}$	2^{4}	2^{2}	$2^{10.8}$
WMC Ratio ID=9	$2^{0.35}$	2^{6}	2^{-10}	$2^{4.8}$
TENET-WH				
WMC Ratio ID=1	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=2	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=3	2^{-10}	2^{-10}	2^{-10}	2^{-10}
WMC Ratio ID=4	2^{10}	2^{-10}	2^{-10}	$2^{4.4}$
WMC Ratio ID=5 (TENET-W)	2^{6}	2^{8}	2^{-10}	$2^{7.6}$
WMC Ratio ID=6	$2^{-2.56}$	$2^{-5.6}$	2^{-10}	$2^{-0.32}$
WMC Ratio ID=7	$2^{-1.36}$	$2^{-3.92}$	2^{-10}	2^{0}
WMC Ratio ID=8	$2^{2.4}$	2^{0}	2^{2}	2^{8}
WMC Ratio ID=9	2^{-10}	2^{-10}	2^{-10}	2^{12}

Table 17: Best model C parameter for the various combined signaling network variants (C_1 to C_4) using different approaches with linear kernel.

ranks of all nodes in the test sets of I_1, I_2 and I_4, respectively. For the remaining networks, due to the larger size of the test sets, we only plot the *normalized ranks* of selected nodes in the test sets in Figure 22 and Figures 23 to 26. The *normalized rank* of a node u for a particular approach x is denoted as $\Psi_{norm(x):u}$ and defined as follows.

$$\Psi_{norm(x):u} = \frac{\Psi_{x:u}}{\text{MAX}_{i \in V}(\Psi_x : i)} \tag{4}$$

where $\Psi_{x:u}$ is the rank of u based on x, V is the set of nodes in the given signaling network and $\text{MAX}(\cdot)$ is the maximum operator. We use the normalized rank for comparison since the range of ranks for each approach is different. In these figures,

	$\mathbf{I}_1$	$\mathbf{I}_2$	$\mathbf{I}_3$
Best Approaches	WRE (C^+=0.9)	WRE (C^+=0.8)	WRE (C^+=0.4)
$\mathcal{P}$	4.843	2.798	3.097
$\overline{\phi}(val)$	0.843	0.898	0.688
$\phi(test)$	1	0.9	0.672
TPR	1	0	0.095
TNR	1	1	0.975
PPV	1	0	0.667
SVM cost parameter C	2^6	2^{-10}	2^{12}

Table 18: Summary of best DIFFER characterization model for different networks.

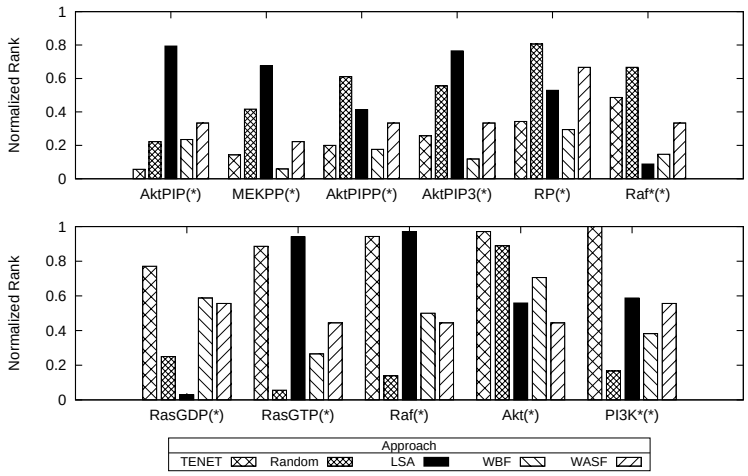

Figure 23: Normalized rank of nodes in test set of C_1.

we mark the known targets in the figures using $(*)$ for ease of reading. The DIFFER approach refers to the method used to find discriminative topological features (DTF) in our previous work [16]. Specifically, we perform SVM using WMC and WRE. We use the WRE feature selection approach to ensure that the DTFs are used specifically for training the SVM and tested the entire range of WMC (*i.e.*, [0–1]). The characterization model with the best performance score $\mathcal{P}$ is then used to generate the prioritization ranks of DIFFER. Table 18 provides a summary of the best characterization models of DIFFER. Compared to TENET, the performance scores of DIFFER's characterization models are consistently lower. Note that an ideal ranking approach should assign higher normalized ranks, represented by shorter histogram, to known targets. From these figures, we note that normalized ranks of a given node can vary widely using different prioritization approaches. For instance, in Figure 24, the normalized rank of `G3P`, a known target, is relatively high when prioritized using TENET, random prioritization and LSA. However, it is given relatively low ranking by WBF and WASF. In another network C_1 (Figure 23), WBF and WASF assigned higher ranks to `Raf*`, another known target, compared to TENET and random prioritization. Hence, an approach that performs better for one particular network can perform poorly in another.

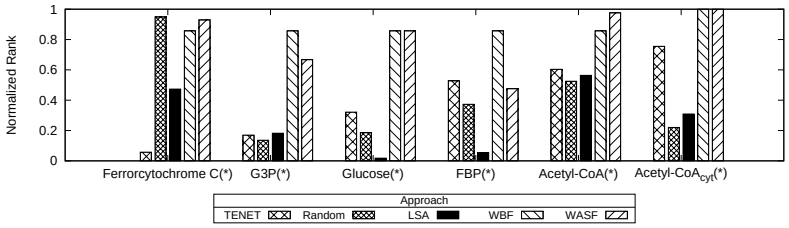

Figure 24: Normalized rank of nodes in test set of C_2.

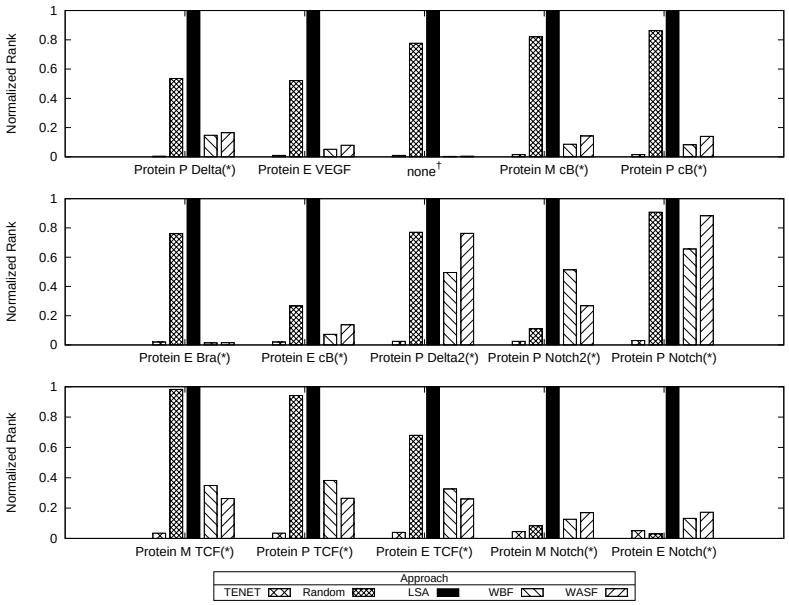

Figure 25: Normalized rank of top 15 nodes ranked using TENET in test set of C_3. †*none* is the node used to represent degraded proteins in this network.

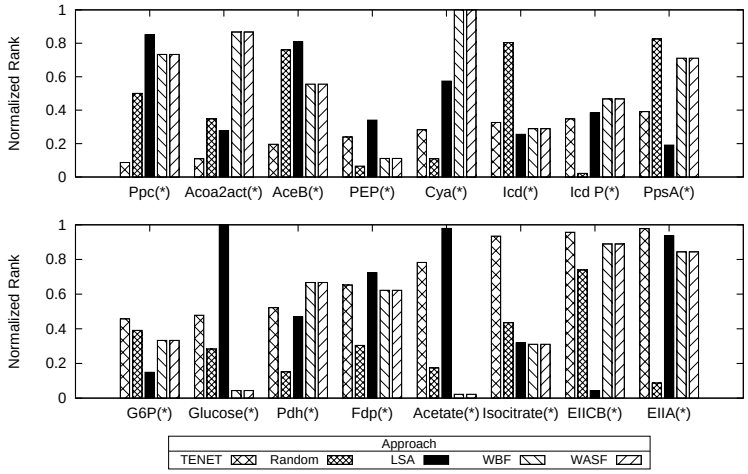

Figure 26: Normalized rank of nodes in test set of C_4.

Next, we discuss the ability of each method to predict known targets in the test set of two networks (I_1 and I_4) in particular. For a given network $G = (V, E)$, containing a set of known targets $T \subset V$, an ideal target prioritization approach should rank all targets higher[19] than non-targets. That is, the set of targets should be contained within the top-($\frac{|T|}{|V|} \times 100$)% ranks of the network. For clarity, a target $t \in V$ is *correctly* predicted if it is given a rank within the top-($\frac{|T|}{|V|} \times 100$)%. In the `MAPK/PI3K` network (I_1), all approaches except LSA[20] correctly predicted the known target (`MEKPP`) (Figure 19). In another network (`glucose` metabolism network, I_4) (Figure 21), TENET performed the best as it correctly predicted the two known targets (`EIIA` and `isocitrate`). In contrast, random prioritization and WASF correctly predicted `EIIA` but missed `isocitrate` (ranked third and fourth in random prioritization and WASF, respectively). The remaining approaches performed worse and missed both targets. LSA and WBF ranked `EIIA` and `isocitrate` as second and fifth, respectively whereas DIFFER did not yield any results as no discriminative topological features (DTF) were found for this network. Note that the ROC AUC is generally used to compare the performance of different characterization models and the results of the ROC AUC can be found in Figure 16. From this figure, we observe that TENET outperforms other approaches in terms of the quality of the prioritization results, particularly for individual networks, and is comparable in terms of runtime performance when SVM training is performed offline (TENET (Regression only)).

4.7 Scalability

Although existing signaling networks tend to be small (tens to hundreds of nodes) in size, they are expected to grow. We tested TENET to assess its scalability to larger networks. The largest curated signaling network we obtain from Biomodels.net was I_3, `endomesoderm` gene regulatory network with over 600 nodes (previous experiments). We further tested TENET on a larger network with 2635 nodes and 43735 edges. This network is obtained from a human signaling network [58] curated by Edwin and colleagues and is implicated in cancer. The targets (324 nodes, 12.3% of dataset) in this network that are deemed relevant to cancer are reported in [17]. We performed 10-fold cross validations and the test set (239 nodes) contains 27 targets (11.3% of test set). TENET took 2 days for the analysis using the following configuration: C parameter range=$[2^{-8} - 2^{8}]$, *libsvm* shrinking parameter=1, *libsvm* SVM type=NU_SVC, C^{+}=0.7. Note that we use this SVM configuration to avoid numeric instability when performing SVM training for the larger network. The weighted misclassification cost was set to 0.7 as larger values in the range of [0.6 - 0.8] were found to perform better (Table 5 in main text). The best SVM model has C parameter 2^8 and has associated predictive features $\mathcal{X}_{all} \backslash \theta_{in}$. The need for multiple predictive features may be due to the presence

[19] Note that in real situations, it is unlikely for us to know the exact ranks within the set of targets.

[20] In LSA, `MEKPP` is ranked third out of four nodes in the test set.

Approach	ROC AUC	AUPR
TENET	0.718	0.251
Random	0.54	0.127
DIFFER	0.627	0.215

Table 19: Comparison of different approaches on a large network.

of different types of cancer targets (*e.g.*, different hallmarks) which may be characterized differently. Hence, we performed further experiments by classifying the targets based on 8 different hallmarks suggested in [35] and characterizing each hallmark category. The hallmark targets were curated using information from the OMIM database [34]. The results from hallmark-based characterization are similar to our previous results. That is, predictive features remain as $\mathcal{X}_{all} \backslash \theta_{in}$ regardless of hallmarks. This further highlights the need of using multiple predictive features to characterize targets and a single feature (*e.g.*, bridging centrality) may not be effective. In addition, we note that certain topological features (*e.g.*, closeness centrality) we study requires the computation of the shortest path which has O($|V|^3$) time complexity using Floyd-Warshall algorithm [29] where $|V|$ is the size of the network. Hence, this may impose an upper limit on the size of the network that TENET can handle. We can address this limitation by extending TENET with techniques that estimate shortest path for large networks such as [81] and [110].

Finally, we compared TENET to other state-of-the-art signaling network-based approaches. Note that comparison could not be performed on LSA and *NetworkPrioritizer*. For the former, no dynamic information of the large cancer network was available whereas for the latter, the program did not complete the analysis due to memory issues. We observe that TENET (ROC AUC=0.718 and AUPR=0.251) outperforms other approaches in terms of ROC AUC and AUPR (Table 19).

5 Conclusions & Future Work

We propose TENET, a SVM-based approach that characterizes known targets in signaling networks using topological features by identifying a set of predictive topological features and using them to generate a characterization model. TENET uses feature selection to remove redundant features, thereby improving prediction accuracy of the characterization models and WMC to improve other performance criteria (*e.g.*, sensitivity). Our empirical study reveals that the characterization models generated by TENET outperforms state-of-the-art approaches in prioritizing signaling and PPI networks. In summary, the contribution of this work is a machine learning-based framework that affords flexibility in characterizing signaling networks of different sizes and with different number of known targets. Although TENET is evaluated on a small[21] number of signaling networks, it can easily incorporate additional signaling networks without any modification to the framework. As part of future work, we intend to explore how the characterization models learnt by

[21] Manual target curation, a time-intensive process, is needed to identify known targets of signaling networks for validating our experimental results.

TENET can be leveraged for target prioritization of signaling networks with *unknown* targets.

References

[1] Albert, R. (2005). Scale-free networks in cell biology. *J. Cell Sci.*, **118**(21), 4947-4957.

[2] Arteaga, C. (2011). The phosphatidylinositol-3 kinase/mTOR pathway: new agents. *Breast Cancer Res.*, **13**(Suppl 2), O8.

[3] Bast, R. (2008). The biology of ovarian cancer: new opportunities for translation. *Nat. Rev. Cancer*, **9**(6), 415-428.

[4] Bhuyan, A. *et al.* (2001). Resting membrane potential as a marker of apoptosis: studies on Xenopus oocytes microinjected with cytochrome c. *Cell Death Differ.*, **8**(1), 63.

[5] Bonacich, P. *et al.* (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, **23**(3), 191-201.

[6] Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. *J. Math. Sociol.*, **25**, 163-177.

[7] Campbell, P. *et al.* (2010). TLN-4601 suppresses growth and induces apoptosis of pancreatic carcinoma cells through inhibition of Ras-ERK MAPK signaling. *J. Mol. Signal.*, **5**(1), 18.

[8] Chang, C.-C. *et al.* (2011). Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**(3), 27.

[9] Chang, D.-E. *et al.* (1999). Acetate metabolism in a pta mutant of Escherichia coli W3110: importance of maintaining acetyl coenzyme A flux for growth and survival. *J. Bacteriol.*, **181**(21), 6656-6663.

[10] Chao, Y.-P. *et al.* (1993). Alteration of growth yield by overexpression of phosphoenolpyruvate carboxylase and phosphoenolpyruvate carboxykinase in Escherichia coli. *Appl. Environ. Microbiol.*, **59**(12), 4261-4265.

[11] Chapelle, O. *et al.* (2002). Choosing multiple parameters for support vector machines. *Mach. Learn.*, **46**(1-3), 131-159.

[12] Chen, J. *et al.* (2009). ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37**(suppl 2), W305-W311.

[13] Chen, L. *et al.* (2005). Stack-based algorithms for pattern matching on DAGs. In *VLDB*.

[14] Chen, Y.-A. *et al.* (2011). Targetmine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PLoS One*, **6**(3), e17844.

[15] Chua, H. *et al.* (2011). PANI: A novel algorithm for fast discovery of putative target nodes in signaling networks. In *ACM BCB*.

[16] Chua, H. *et al.* (2014). One feature doesn't fit all: Characterizing topological features of targets in signaling networks. In *ACM BCB*.

[17] Cui, Q. *et al.* (2007). A map of human cancer signaling. *Mol. Syst. Biol.*, **3**(152).

[18] Cunningham, C. *et al.* (2000). A phase I trial of c-Raf kinase antisense oligonucleotide ISIS 5132 administered as a continuous intravenous infusion in patients with advanced cancer. *Clin. Cancer Res.*, **6**(5), 1626-1631.

[19] Deeb, D. *et al.* (2010). Growth inhibitory and apoptosis-inducing effects of xanthohumol, a prenylated chalone present in hops, in human prostate cancer cells. *Anticancer Res.*, **30**(9), 3333-3339.

[20] Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, **1**(1), 269-271.

[21] Ding, B. *et al.* (2011). Fast set intersection in memory. In *VLDB*.

[22] Emanuele, N. *et al.* (1998). Consequences of alcohol use in diabetics. *Alcohol Health Res. W.*, **22**, 211-219.

[23] Engelfiet, J. *et al.* (1990). A comparison of boundary graph grammars and context-free hypergraph grammars. *Inform. Comput.*, **84**(2), 163-206.

[24] Erion, M. *et al.* (2005). MB06322 (CS-917): a potent and selective inhibitor of fructose 1, 6-bisphosphatase for controlling gluconeogenesis in type 2 diabetes. *PNAS*, **102**(22), 7970-7975.

[25] Espinosa, I. *et al.* (2010). Tagatose: from a sweetener to a new diabetic medication? *Expert Opin. Investig. Drugs*, **19**(2), 285-294.

[26] Fagiolo, G. (2007). Clustering in complex directed networks. *Phys. Rev. E*, **76**(2), 026107-026114.

[27] Farmer, W. *et al.* (1997). Reduction of aerobic acetate production by Escherichia coli. *Appl. Environ. Microbiol.*, **63**(8), 3205-3210.

[28] Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Mach. Learn.*, 31, 1-38.

[29] Floyd, R. (1962). Algorithm 97: shortest path. *Comm. ACM*, **5**(6), 345.

[30] Fontecilla-Camps, J. *et al.* (2009). Structure-function relationships of anaerobic gas-processing metalloenzymes. *Nature*, **460**(7527), 814-822.

[31] Fredman, M. *et al.* (1987). Fibonacci heaps and their uses in improved network optimization algorithms. *JACM*, **34**(3), 596-615.

[32] Freeman, L. (1979). Centrality in social networks conceptual clarification. *Social Networks*, **1**(3), 215-239.

[33] Gustafson, P. *et al.* (1996). Local sensitivity analysis. *Bayesian statistics*, **5**, 197-210.

[34] Hamosh, A. *et al.* (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**(suppl 1), D514-D517.

[35] Hanahan, D. *et al.* (2000). The hallmarks of cancer. *Cell*, **100**(1), 57-70.

[36] Hatakeyama, M. *et al.* (2003). A computational model on the modulation of mitogen-activated protein kinase (mapk) and akt pathways in heregulin-induced erbb signalling. *Biochem. J.*, **373**(Pt 2), 451-463.

[37] He, X. *et al.* (2006). Why do hubs tend to be essential in protein networks? *PLoS Genet.*, **2**(6), e88.

[38] Hosmer Jr., D. *et al.* (2004). *Applied Logistic Regression. Second Edition.* John Wiley & Sons.

[39] Hsu, C.-W. *et al.* (2003). A practical guide to support vector classification.

[40] Hu, E. *et al.* (2012). White rice consumption and risk of type 2 diabetes: meta-analysis and systematic review. *BMJ*, **344**, e1454.

[41] Hwang, W.-C. *et al.* (2008). Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery. *Clin. Pharma. Ther.*, **84**(5), 563-572.

[42] Infante, J. *et al.* (2010). Safety and efficacy results from the first-in-human study of the oral MEK 1/2 inhibitor GSK1120212. *J. Clin. Oncol.*, **28**(15), 2503.

[43] Jahn, S. *et al.* (2013). A role for EIIANtr in controlling fluxes in the central metabolism of E. coli K12. *Biochim. Biophys. Acta,*, **1833**(12), 2879-2889.

[44] Jiang, N. *et al.* (2007). A kinetic core model of the glucose-stimulated insulin secretion network of pancreatic β cells. *Mamm. Genome*, **18**(6-7), 508-520.

[45] Johnston, C. *et al.* (2006). Vinegar: medicinal uses and antiglycemic effect. *Med. Gen. Med.*, **8**(2), 61.

[46] Kabir, M. *et al.* (2004). Metabolic regulation analysis of icd-gene knockout Escherichia coli based on 2D electrophoresis with MALDI-TOF mass spectrometry and enzyme activity measurements. *Appl. Microbiol. Biot.*, **65**(1), 84-96.

[47] Kacprowski, T. *et al.* (2013). Networkprioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, **29**(11).

[48] Kang, U. *et al.* (2011). Centralities in large networks: algorithms and observations. In *SDM*.

[49] Kiss, C. *et al.* (2008). Identification of influencersmeasuring influence in customer networks. *Decis. Suppor Syst.*, **46**(1), 233-253.

[50] Klamt, S. *et al.* (2009). Hypergraphs and cellular networks. *PLoS Comput. Biol.*, **5**(5), e1000385.

[51] Klip, A. *et al.* (1990). Cellular mechanism of action of metformin. *Diabetes Care*, **13**(6), 696-704.

[52] Kohavi, R. *et al.* (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI*, **14**, 1137-1145.

[53] Kohl, C. *et al.* (2002). Effects of benfluorex on fatty acid and glucose metabolism in isolated rat hepatocytes: from metabolic fluxes to gene expression. *Diabetes*, **51**(8), 2363-2368.

[54] Kondapaka, S. *et al.* (2003). Perifosine, a novel alkylphospholipid, inhibits protein kinase B activation. *Mol. Cancer Ther.*, **2**(11), 1093-1103.

[55] Kotte, O. *et al.* (2010). Bacterial adaptation through distributed sensing of metabolic fluxes. *Mol. Syst. Bio.*, **6**(1), 355.

[56] Kwon, Y.-K. *et al.* (2008). Coherent coupling of feedback loops: a design principle of cell signaling networks. *Bioinformatics*, **24**(17), 1926-1932.

[57] Lavigne, C. *et al.* (2001). Prevention of skeletal muscle insulin resistance by dietary cod protein in high fat-fed rats. *Am. J. Physiol. Endocrinol. Metab.*, **281**(1), E62-E71.

[58] Li, L. *et al.* (2002). The human phosphotyrosine signaling network: evolution and hotspots of hijacking in cancer. *Genome Res.*, **22**(7), 1222-1230.

[59] Li, T. *et al.* (2001). Performance of batch membrane reactor: Glycerol-3-phosphate synthesis coupled with adenosine triphosphate regeneration. *Biotechnol. Bioeng.*, **74**(4), 326-334.

[60] Longabaugh, W. (2012). BioTapestry: a tool to visualize the dynamic properties of gene regulatory networks. *Gene Regulatory Networks*, Springer, 359-394.

[61] Machicao, F. *et al.* (2012). Pleiotropic neuroprotective and metabolic effects of actovegin's mode of action. *J. Neurol. Sci.*, **322**, 222-227.

[62] Maira, S.-M. *et al.* (2008). Identification and characterization of nvp-bez235, a new orally available dual phosphatidylinositol 3-kinase/mammalian target of rapamycin inhibitor with potent in vivo antitumor activity. *Mol. Cancer Ther.*, **7**(7), 1851-1863.

[63] Marill, T. *et al.* (1963). On the effectiveness of receptors in recognition systems. *IEEE Trans. Inf. Theory*, **9**(1), 11-17.

[64] Martínez, V. *et al.* (2012). Network-based gene-disease prioritization using PROPHNET. *EMBnet. J.*, **18**(B), 38.

[65] McDermott, J. *et al.* (2012). Topological analysis of protein co-abundance networks identifies novel host targets important for hcv infection and pathogenesis. *BMC Syst. Biol.*, **6**(1), 28.

[66] Meyer, D. (2012). Support vector machines. *The Interface to libsvm in package e1071*. *e1071 Vignette*.

[67] Mishra, R. *et al.* (2011). Jiao gu lan (gynostemma pentaphyllum): The chinese rasayan-current research scenario. *Int. J. Res. Pharm. Biomed. Sci.*, **2**(4), 1483.

[68] Morgan, S. *et al.* (2011). The cost of drug development: a systematic review. *Health Policy*, **100**(1), 4-17.

[69] Murarka, A. *et al.* (2010). Metabolic analysis of wild-type Escherichia coli and a pyruvate dehydrogenase complex (PDHC)-deficient derivative reveals the role of PDHC in the fermentative metabolism of glucose. *J. Biol. Chem.*, **285**(41), 31548-31558.

[70] Nagarajan, N. *et al.* (2009). Reliability and efficiency of algorithms for computing the significance of the Mann-Whitney test. *Comput. Stat.*, **24**(4), 605-622.

[71] Newman, M. (2006). The mathematics of networks.

[72] National Cancer Institute. NCI drug dictionary. *http://www.cancer.gov/drugdictionary/*, Accessed 5 May 2014.

[73] NIH. *http://www.clinicaltrials.gov*, Accessed 5 May 2014.

[74] Patnaik, R. *et al.* (1992). Stimulation of glucose catabolism in Escherichia coli by a potential futile cycle. *J. Bacteriol.*, **174**(23), 7527-7532.

[75] Pavlopoulos, G. *et al.* (2011). Using graph theory to analyze biological networks. *BioData Min.*, **4**(1), 1-27.

[76] Perrenoud, A. *et al.* (2005). Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in Escherichia coli. *J. Bacteriol.*, **187**(9), 3171-3179.

[77] Phue, J.-N. *et al.* (2004). Transcription levels of key metabolic genes are the cause for different glucose utilization pathways in E. coli B (BL21) and E. coli K (JM109). *J. Biotechnol.*, **109**(1-2), 21-30.

[78] Phue, J.-N. *et al.* (2005). Glucose metabolism at high density growth of E. coli B and E. coli K: Differences in metabolic pathways are responsible for efficient glucose utilization in E. coli B as determined by microarrays and Northern blot analyses. *Biotechnol. Bioeng.*, **90**(7), 805-820.

[79] Picon, A. *et al.* (2005). Reducing the glucose uptake rate in Escherichia coli affects growth rate but not protein production. *Biotechnol. Bioeng.*, **90**(2), 191-200.

[80] Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods*, 185-208.

[81] Potamias, M. (2009). Fast shortest path distance estimation in large networks. In *CIKM*.

[82] Qiu, C. *et al.* (2008). Mechanism of activation and inhibition of the HER4/ErbB4 kinase. *Structure*, **16**(3), 460-467.

[83] RxList Inc. RxList-The internet drug index. *http://www.rxlist.com*. Accessed 3 September 2013.

[84] Saeys, Y. *et al.* (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**(19), 2507-2517.

[85] Sahle, S. *et al.* (2006). Simulation of biochemical networks using copasi: a complex pathway simulator. In *WSC*, 1698-1706.

[86] Scardoni, G. *et al.* (2012). New Frontiers in Graph Theory. Chapter 16.

[87] Shiloach, J. *et al.* (2009). Glucose and acetate metabolism in E. coli–system level analysis and biotechnological applications in protein production processes. *Systems biology and biotechnology of Escherichia coli*, Springer, 377-400.

[88] Spencer, A. *et al.* (2013). Novel AKT inhibitor afuresertib in combination with bortezomib and dexamethasone demonstrates favorable safety profile and significant clinical activity in patients with relapsed/refractory multiple myeloma. *Blood*, **122**(21), 283.

[89] Stark, C. *et al.* (2011). The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**(suppl 1), D698-D704.

[90] Steriti, R. (2010). Berberine for diabetes mellitus type 2. *Natural Medicine J.*, **2**(10), 1-5.

[91] Takes, F. *et al.* (2013). Computing the eccentricity distribution of large graphs. *Algorithms*, **6**(1), 100-118.

[92] Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM J. Sci. Comput.*, **1**(2), 146-160.

[93] Tsang, I. *et al.* (2005). Core vector machines: fast SVM training on very large data sets. *J. Mach. Learn. Res.*, **6**(4), 363-392.

[94] van de Walle, M. *et al.* (1998). Proposed mechanism of acetate accumulation in two recombinant escherichia coli strains during high density fermentation. *Biotechnol. Bioeng.*, **57**(1), 71-78.

[95] van Poelje, P. *et al.* (2011). Fructose-1,6-bisphosphatase inhibitors for reducing excessive endogenous glucose production in type 2 diabetes. *Diabetes-Perspectives in Drug Therapy*, 279-301.

[96] Venkatesan, A. *et al.* (2010). Bis(morpholino-1,3,5-triazine) derivatives: Potent adenosine 5'-triphosphate competitive phosphatidylinositol-3-kinase/mammalian target of rapamycin inhibitors: discovery of compound 26 (PKI-587), a highly efficacious dual inhibitor. *J. Med. Chem.*, **53**(6), 2636-2645.

[97] Verspohl, E. (2012). Novel pharmacological approaches to the treatment of type 2 diabetes. *Pharmacol. Rev.*, **64**(2), 188-237.

[98] Viollet, B. *et al.* (2012). Cellular and molecular mechanisms of metformin: an overview. *Clin. Sci.*, **122**(6), 253-270.

[99] Watts, D. *et al.* (1998). Collective dynamics of śmall-worldńetworks. *Nature*, **393**(6684), 440-442.

[100] Wilhelm, S. *et al.* (2006). Discovery and development of sorafenib: a multikinase inhibitor for treating cancer. *Nat. Rev. Drug Discov.*, **5**(10), 835-844.

[101] Williams, V. (2012). Multiplying matrices faster than Coppersmith-Winograd. In *STOC*.

[102] Wishart, D. *et al.* (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**(Database issue), D901-D906.

[103] Wuchty, S. *et al.* (2003). Centers of complex networks. *J. Theor. Biol.*, **223**(1), 45-53.

[104] Yan, X. *et al.* (2013). The identification of novel targets of mir-16 and characterization of their biological functions in cancer cells. *Mol. Cancer*, **12**, 92.

[105] Yang, K.*et al.* (2008). Finding Multiple Target Optimal Intervention in Disease-related Molecular Network. *Mol. Syst. Biol.*, vol. 4.

[106] Yap, T.*et al.* (2011). First-in-man clinical trial of the oral pan-AKT inhibitor MK-2206 in patients with advanced solid tumors. *J. Clin. Oncol.*, **29**(35), 4688-4695.

[107] Yeh, T.*et al.* (2007). Biological characterization of ARRY-142886 (AZD6244), a potent, highly selective mitogen-activated protein kinase kinase 1/2 inhibitor. *Clin. Cancer Res.*, **13**(5), 1576-1583.

[108] Zhang, J. *et al.* (2010). Novel biological network features discovery for in silico identification of drug targets. In *IHI*, 144-152.

[109] Zhao, L. *et al.* (2004). Evaluation of combination chemotherapy. *Clin. Cancer Res.*, **10**(23), 7994-8004.

[110] Zhao, X. *et al.* (2010). Orion: shortest path estimation for large social graphs. *Networks*, **1**, 5.