TheraNet: An Enhanced in Silico Network-Based Drug Efficacy Screening with a Novel Approach for Weighing Candidate Drugs and Disease Genes

Introduction

The advent of network-based methodologies has revolutionized the landscape of drug discovery by providing a more comprehensive view of biological systems. Traditional approaches to drug efficacy screening often focus on individual molecular targets, which can be limited in scope and overlook the broader complexity of biological networks. In contrast, network-based models integrate information about the interactions between genes, proteins, and other biomolecules, enabling a systems-level approach to identifying potential drug candidates. One such method, detailed in the paper "Network-based in silico drug efficacy screening," utilizes a computational framework to prioritize drugs based on their ability to modulate disease-associated genes through network connectivity.

While this approach has garnered significant attention due to its innovative use of biological networks, a closer analysis of the methodology revealed critical errors that could affect the robustness and reliability of its results. Specifically, inaccuracies in how network connections were weighted and how drug-disease associations were evaluated raised concerns about the method's capacity to accurately prioritize candidate drugs. Given the potential of this approach to streamline drug repurposing efforts and improve treatment discovery, it is essential to address these limitations and refine the existing framework.

In this study, we reanalyze the original methodology, identifying and correcting key errors that influence the network's performance. The improvements we introduce therapeutic network-based drug efficacy weighing method (TheraNet) ensure a more accurate representation of the interactions between drugs and disease-associated genes, refining the predictive capabilities of the model. In addition to these corrections, we introduce a novel weighting scheme that assigns dynamic weights to both drugs and disease genes based on their biological significance and network properties. This approach enhances the model's ability to account for the varying influence of different genes within a disease network and the relative efficacy of drugs targeting those genes.

By integrating this enhanced weighting mechanism, our method not only addresses the issues found in the original approach but also pushes the boundaries of in silico drug screening, making it more precise and biologically relevant. The findings of this study hold significant implications for drug discovery, particularly in the context of drug repurposing, where the need for rapid identification of therapeutic agents is paramount. We believe that the updated method will provide researchers with a more powerful tool to predict drug efficacy and accelerate the development of treatments for complex diseases, ultimately contributing to more effective and personalized healthcare solutions.

Method and Discussion

Benefits of the Network-Based in Silico Drug Efficacy Screening Approach

The network-based in silico drug efficacy screening method offers distinct advantages over traditional drug discovery approaches by utilizing a systems-level perspective. Instead of focusing on individual molecular targets, it analyzes complex interactions between genes, proteins, and drugs within a biological network, providing a more comprehensive understanding of disease mechanisms.

One major benefit is its ability to facilitate drug repurposing by identifying existing drugs that can target disease-associated genes. This significantly reduces the time and cost of drug development. Additionally, the method efficiently prioritizes candidate drugs based on their network connectivity, allowing researchers to focus on the most promising candidates and optimize resources. The method's flexibility is enhanced by its ability to integrate multi-omics data, improving the accuracy of drug predictions across diverse biological contexts. It is also disease-agnostic, making it applicable to a wide range of diseases without needing specific adjustments. Another strength is its potential to reveal insights into the mechanisms of drug action by analyzing interactions within the network, potentially uncovering unknown drug effects. Finally, this in silico framework enables data-driven hypotheses, streamlining experimental validation by highlighting the most relevant drug-disease associations.

In summary, this network-based method accelerates drug discovery by providing a comprehensive, flexible, and efficient approach to identifying and prioritizing candidate drugs, especially in complex diseases.

hypothesis to investigate the relationship between drug targets and disease proteins

The paper hat five hypothesis to investigate the relationship between drug targets and disease proteins.

1. **Closest (d_c)**: Measures the **average shortest path length** between each target of a drug and its **closest disease protein** in the network. If a drug target is close to any disease protein, it is more likely to have an impact. This method focuses on the **nearest connection** for each target.

2. **Shortest (d_s)**: Computes the **average of all shortest path lengths** between **all drug targets** and **all disease proteins**. **Interpretation**: Rather than focusing on the closest connection, this method evaluates the overall connectivity between the drug targets and the disease proteins. It captures a **global relationship** between the two sets.

3. **Kernel (d_k)**: Uses a **graph kernel** to calculate the similarity between the drug targets and the disease proteins. It models the spread of influence or similarity between nodes based on the network structure. Incorporates **global network properties** but can be computationally expensive and may overfit when the network is too dense or sparse.

4. **Center (d_{cc})**: Evaluates the **centrality** of drug targets in the disease module by measuring how close the drug targets are to the center of the disease module. The assumption is that if drug targets are highly central within the disease network, they might disrupt disease progression more effectively.

5. **Separation (d_{ss})**: Calculates the **topological separation** between drug targets and the disease module. It considers whether the targets are **close to the disease proteins** or **well-separated** from them in the network. This measures whether the drug targets form their own module (independent of the disease module) or interact with the disease module.

The paper "Network-based in silico drug efficacy screening" concludes that shortest (d_s) is the most accurate method for predicting drug efficacy because: (F**igure 1**)

1. **Comprehensive View**: While methods like $\mathbf{d_c}$ only consider the nearest connections, $\mathbf{d_s}$ takes into account the overall connectivity between all drug targets and all disease proteins. This avoids biases introduced by focusing on a single closest interaction.
2. **Captures Global Network Features:** Unlike methods that rely on centrality ($\mathbf{d_{cc}}$) or modularity ($\mathbf{d_{ss}}$), $\mathbf{d_s}$ reflects the general relationship between drug and disease modules, ensuring that both local and global network properties are captured.
3. **Reduces Noise from Outliers**: Methods like $\mathbf{d_c}$ can be heavily influenced by individual outliers (e.g., one very close drug target). $\mathbf{d_s}$, by averaging over all pairwise distances, smoothens out such noise, making the prediction more robust.

4. **Empirical Validation**: When tested against experimental drug efficacy data, $\mathbf{d_s}$ consistently outperformed other metrics, demonstrating better predictive power across various datasets.

While shortest (d_s) is the most accurate method for predicting drug efficacy, we choose it as a network-based method for drug discovery for the diseases need new drugs.

Network-based proximity between drugs and diseases

The paper "Network-based in silico drug efficacy screening" utilizes protein-protein interaction (PPI) networks to predict the efficacy of drugs by analyzing interactions between drug-target proteins and disease-associated proteins. By constructing a PPI network, the authors identify the relationships between drug targets and disease proteins, aiming to understand how drugs interact with disease mechanisms at a molecular level. The network-based approach evaluates various topological metrics, such as the shortest path length, kernel similarity, and centrality measures, to determine the strength and relevance of drug-target interactions. This enables the identification of promising drug candidates by assessing their proximity and connectivity to disease proteins within the network, thereby enhancing drug efficacy prediction through computational means.

Protein-Protein Interaction (PPI) refers to the physical or functional interactions between proteins, which are crucial for various biological processes within a cell. PPIs are central to cellular functions, including signal transduction, immune responses, and metabolic pathways. In computational biology, PPIs are typically represented as networks, where proteins are nodes and interactions are edges connecting them. To calculate PPIs, various experimental methods like yeast two-hybrid screens, co-immunoprecipitation, or mass spectrometry are employed to identify direct interactions. Additionally, computational approaches predict PPIs based on sequence or structural similarities, protein domains, or evolutionary data. These interactions are often quantified using metrics such as interaction score or probability to represent the strength or reliability of the PPI, which can be used to build PPI networks that help understand protein function and disease mechanisms.

The STRING database scores protein-protein interactions (PPIs) using a combination of evidence sources, including experimental data, computational predictions, and known interactions from other databases. The interaction score, called the confidence score, ranges from 0 to 1, with higher values indicating stronger evidence for the interaction. This score integrates data from various methods such as co-expression, co-localization, and text mining, providing a robust measure of the likelihood that two proteins interact biologically. The confidence score helps prioritize PPIs for further analysis or experimental validation.

In the STRING database, a confidence score of 0.7 or higher is typically considered to indicate a highly reliable protein-protein interaction (PPI). Interactions with scores above 0.7 are considered well-supported by multiple lines of evidence, such as experimental data or high-confidence computational predictions. These interactions are more likely to represent biologically relevant

connections between proteins, while scores below 0.7 suggest weaker evidence or less confidence in the interaction. Therefore, in Guney's paper and our study, a confidence score of 0.7 is used as a threshold to classify protein-protein interactions (PPIs) as either valid or invalid. Interactions with a score of 0.7 or higher are considered valid and are included in the analysis, as they are supported by reliable evidence from experimental data or computational predictions. Interactions with scores below 0.7 are considered less reliable and are excluded from the PPI network, as the evidence supporting these interactions is not strong enough to confidently infer a biological connection between the proteins.

Identifying and Resolving Issues in the Mean Distance Algorithm

Drug gene interaction scores

DGIdb 5.0 stands as a cornerstone in the field of precision medicine and drug discovery, offering a meticulously curated database that maps the interactions between drugs and genes. This rich dataset, compiled from diverse sources, empowers researchers to explore critical connections that can fuel the development of novel therapies and the identification of personalized treatment strategies. The importance of DGIdb lies in its ability to provide insights that bridge the gap between drug efficacy and genetic variability, thereby enhancing the prospects of tailored, effective treatments for individuals. To maximize the utility of this valuable resource, normalizing the ranks of drug-gene interactions to a scale between 0 and 1 allows for more consistent and comparable data analysis. However, the necessity of omitting low-rank interactions (those smaller than 0.01) is equally critical. By removing these lower-confidence associations, researchers can ensure that their analyses focus on high-quality, biologically significant drug-gene pairs. This step reduces noise and enhances the reliability of downstream findings, ensuring that only the most promising interactions are considered in drug repurposing, biomarker discovery, and other clinical applications, ultimately driving precision medicine forward with greater confidence.

Include proteins coding genes without interaction in the distance calculation

Figure 2.a shows an example of an association between a drug and a disease. Drug targets proteins t_1,t_2, and t_3 and disease targets effects proteins s_1 and s_2 . To calculate d_s , we should find the shortest distance between each drug target protein and disease target protein set. The closest disease target protein to t_1 is s_2, means the shortest path between t_1 and set $\{s_1,s_2,s_3\}$ equals 2. With the same algorithm, closest disease target shortest path between t_2 and set $\{s_1,s_2,s_3\}$ equals 3. while we use shortest (ds) method. The average ((2+3)/2) is the calculated value between drug and disease target proteins. (F**igure 2.b**)

Figure 3 shows the possible association between drug and disease target proteins. It is possible that disease target protein has no interaction with other proteins, but it is a direct target of disease as well (subset A). Then the closest distance between this drug target and disease protein set is zero. On the other hand, If a drug target protein has no interaction with other proteins (**Figure 3, subset E)**, them the shortest distance between that drug target protein and disease target proteins will be infinity. This one increases the average of shortest distance(d_s) significantly. In the

implemented algorithm by Guney, drug proteins with no PPI (F**igure 3. subset E**) and disease target proteins with no PPI (F**igure.3 subset F**) are excluded from the network, and it causes false negative and false positive discoveries.

Figure 4 shows a toy Protein-Protein Interaction network. **Figure 4.a** shows two set of drug target proteins $T=\{t_1,t_2,t_3\}$ and disease target proteins $S=\{s_1,s_2,s_3\}$, and all nodes from set T are connected to at least one of disease proteins in set S. As a result, the average shortest distance between sets T and S is 2. In **Figure 4.b** the drug target proteins $T=\{t_1,t_2,t_3,t_4\}$, and disease targets proteins $S=\{s_1,s_2,s_3\}$. While the disease target protein t_4 does not interaction with any disease target proteins, the shortest distance between t_4 and disease targets proteins is infinity. If we assume infinity value is 100 (it must be big enough in comparison to network size. In this toy example 100 equals infinity, but in the study is value equals 20,000). The average is 26.5. **Figure 4.c** shows a drug targets $T=\{t_1,t_2,t_3,t_4\}$ and a disease targets proteins $S=\{s_1,s_2,s_3,s_4\}$. The protein t_4 is the same proteins as s_4. It means the shortest stridence between proteins t_4 and disease protein target set, S, is zero. Therefor the average shortest distance between the new drug and disease sets will be 1.5. with this example we see the importance and keeping proteins with no interaction in the network. We fixed this problem in the revised version.

z-score threshold

The selection of an appropriate z-score threshold is crucial for ensuring statistical reliability in drug network analyses. A z-score of -1.96 corresponds to the 5th percentile in a standard normal distribution, which is a well-established significance threshold in hypothesis testing ($p < 0.05$, one-tailed test). This cutoff minimizes false discoveries by ensuring that only statistically meaningful deviations are considered significant. In contrast, the previously used threshold of -0.15 is extremely weak and lacks statistical justification. A z-score of -0.15 corresponds to a p-value of approximately 0.44, meaning that nearly 44% of the observations would be expected by chance alone under the null hypothesis (**Figure 5**). Such a high acceptance rate for insignificant variations severely compromises result reliability, leading to the inclusion of spurious associations that do not reflect true biological interactions. Given that drug networks rely on precise statistical filtering to identify functionally relevant associations, the use of an unjustifiably lenient threshold undermines the validity of conclusions. Thus, adopting a more rigorous threshold of -1.96 aligns with established statistical best practices, ensuring that the reported associations are not artifacts of random variation but instead represent biologically meaningful findings.

Random sampling size

Guney compares the network-based distance d_s between the known drug targets and the disease proteins with randomly selected proteins with same degrees of drugs. The distance d_s between drug proteins and disease targets must be closer than expected for randomly selected protein sets, else that drug does not target the disease better than randomly selected protein sets. Drug *"Rucaparib Camsylate*" targets three genes "*PARP1*", "*PARP2*", and "*PARP3*" with degrees 158, 44, and 14. There are 290,147, and 102 similar genes with the same degree in PPI network. It means there are 290x147x102=4,348,260 sub-networks in the protein-protein interaction network with three nodes

and the same degree size. Therefore, we 1000 sampling is not enough to calculate z-score and p-value for such a huge network and prove it with some evidence. (The pseudocode in **Figure 6** shows the method of p-value and z-score calculation for drug and disease gene set pairs.)

With 1k random drug sampling, the THERANET method suggest 40 drungswith z-score <= -1.96 and p-value <= 0.05, while this value equals 35 for 10k random sampling and 37 for 100k random sampling. The algorism with 1k random sampling, suggest 3 drugs which are not suggested with 100k random sampling. The method with 10k sampling could not detect two drugs which are suggested with 100k random sampling. This comparison shows that z-score and p-values are stable enough with 1k or 10k random drug sampling, and there are some false positive (3 for 1k) or false negative (2 for 10k) results.

To evaluate the robustness of our statistical analysis, we compared the z-scores, p-values, means, and standard deviations derived from three random sampling sizes: 1k, 10k, and 100k samples per drug. We assessed these metrics using four statistical measures: Standard Deviation of Differences (SDD), Mean Difference, Range of Differences, and Coefficient of Variation (CV). The comparison between **1k and 100k samples** revealed substantial discrepancies. The **SDD of 0.227** and a **range of differences of 6.521** in z-scores indicate significant variability when using only 1k samples, suggesting instability in the results. Similarly, the **SDD for mean differences (8.065)** and standard deviations (49.452) underscore the unreliability of smaller sample sizes, with the **CVs of 23.859 for means** and **20.178 for SD** further reflecting high relative variability. These findings demonstrate that 1k samples introduce considerable noise, undermining the precision and consistency of the statistical estimates.

In contrast, the comparison between **10k and 100k samples** showed much smaller discrepancies. The **SDD of 0.011** and a **range of differences of 0.119** for z-scores, alongside a more modest **SDD of 3.038** for means and **16.482** for standard deviations, indicate that 10k samples yield results closely aligned with those from 100k samples. Although 100k samples provide marginally better precision—evident in the reduced variability metrics—the **CVs of 49.488 (mean)** and **8.338 (SD)** suggest that 10k samples are sufficient for producing stable and reliable results in most scenarios. These comparisons highlight that while **1k samples are inadequate** due to high variability and inconsistency, **10k samples provide a reliable balance** of accuracy and computational efficiency. However, **100k samples remain the gold standard** for ensuring maximum precision, especially in sensitive analyses where minimal variability is crucial (**Table 1** shows the details).

The running time increases by factor 1000. Hopefully, the random sampling can be done per each drug separately it means using parallel computing is possible and it can significantly reduce the time for calculation. We repeated the method with 1000 (1k), 10,000 (10k) and 100,000 (100k) random sampling.

Ranking Drugs Based on Drug-Gene and Protein-Protein Interaction (PPI) Scores

Ranking drugs is a critical step in drug discovery and precision medicine, as it helps prioritize compounds with the highest potential for therapeutic efficacy. By systematically evaluating drugs based on their interactions with disease-associated genes and their influence within protein-protein interaction (PPI) networks, researchers can identify the most promising candidates for further development. This ranking process integrates Drug-Gene Interaction (DGI) scores and PPI scores to provide a comprehensive assessment. Each drug target gene receives a rank based on the strength and number of PPI connections it has with disease target genes. Subsequently, the overall drug rank is calculated by combining these PPI-based scores with the DGI scores, ensuring that both direct interactions and broader network effects are considered.

Considering PPI scores between drug target genes and disease target genes offers significant biological benefits. It accounts for the functional relationships and interaction dynamics within the cellular environment, reflecting how perturbations in one part of the network can influence disease pathways. This approach ensures that drugs targeting genes with strong, relevant interactions to disease-associated proteins are prioritized, increasing the likelihood of therapeutic efficacy. Additionally, it captures indirect effects where drugs might modulate disease pathways through intermediate proteins, providing a more holistic view of potential drug action.

The inclusion of Drug-Gene Interaction (DGI) scores is equally important, as it validates the biological relevance and specificity of the drug's action on its target genes. DGI scores quantify the strength and confidence of known interactions between drugs and their target genes, ensuring that only biologically meaningful and experimentally supported interactions contribute to the drug ranking. This dual consideration of PPI and DGI scores enhances the robustness of drug prioritization, facilitating the identification of the most promising therapeutic candidates.

Step 1: Assigning Weights to Disease Target Genes

The weight of each disease-associated gene is determined by the summation of its PPI scores with neighboring genes. This reflects the centrality and potential influence of the gene within the network.

- **Weight of** =
- **Weight of** =

Step 2: Calculating Weights for Drug Target Genes

The weight of each drug target gene is computed by summing the weighted paths from the drug target gene to all disease-associated genes. Each path's weight is the product of the PPI scores of edges in the path and the weight of the corresponding disease-associated gene.

- **Weight of** =

- **Weight of** =

Step 3: Calculating the Overall Drug Score

The final score for drug 'X1' is derived from the sum of the products of drug-gene interaction scores and the corresponding weights of the drug target genes.

- **Weight of** =

Where:

- represents the PPI score of the interaction.
- represents the drug-gene interaction score for the target gene.

$$Weight\ drug\ = \sum_{t \in T} dgi_t * \left(\sum_{\substack{s \epsilon S, \\ p \epsilon P[t \to s]}} \left(\prod_{e \in p} PPI_e * \left(1 + \sum_{n \in N[s]} PPI_n \right) \right) \right)$$

$$\boldsymbol{T} = set\ of\ drug\ target\ genes, like\ \ node\ \boldsymbol{t}$$

$$\boldsymbol{S} = set\ of\ disease\ target\ genes, like\ node\ \boldsymbol{s}$$

$$\boldsymbol{P}[\boldsymbol{t} \to \boldsymbol{s}]$$

$$= set\ of\ paths\ from\ node\ \boldsymbol{t}\ to\ node\ \boldsymbol{s}, like\ \ path\ \boldsymbol{p}, and\ \boldsymbol{e}\ is\ and\ edge\ in\ the\ path\ \boldsymbol{p}$$

$$\boldsymbol{N}[\boldsymbol{s}] = set\ of\ edges\ from\ node\ \boldsymbol{s}\ to\ its\ neighbors, like\ edge\ \boldsymbol{n}$$

Text Mining and NLP for Drug Characterization and Side Effect Analysis

Understanding drug characteristics, including their descriptions, formulations, and potential side effects on specific tissues, is crucial for drug safety assessments and pharmacological studies. This study employs text mining and natural language processing (NLP) techniques to extract structured information from online drug databases, including DrugBank, Drugs.com, and RxList. We aim to analyze a list of drugs, extract their descriptions, classify their formulation types (e.g., cream, tablet, capsule), and assess potential side effects on tissues relevant to our study.

We designed a systematic workflow to achieve these objectives, beginning with data collection from the three specified sources. Where applicable, drug-related textual data were extracted using web scraping and API-based retrieval methods. To standardize and structure the information, the collected text underwent preprocessing steps such as tokenization, stopword removal,

lemmatization, and named entity recognition (NER). Additionally, we employed regular expressions and dictionary-based methods to accurately classify drug formulations.

To identify side effects on specific tissues, we developed a classification system that extracts tissue-related sentences and identifies severity keywords within them. The model categorizes severity into levels such as 'high' and 'medium' by analyzing key phrases related to drug-induced effects on tissues. This approach enables a structured assessment of potential side effects based on predefined patterns and keyword searches, ensuring consistency and accuracy.

Our findings highlight the effectiveness of NLP techniques in extracting structured drug-related information from unstructured online sources. We successfully retrieved drug descriptions, accurately classified their formulations, and identified potential tissue-specific side effects. Future improvements involve refining the keyword-based classification method, expanding the dataset for broader coverage, and integrating advanced validation techniques to enhance precision and reliability.

Results

In section **Method and Discussion,** we saw that there are some issued related to the method and we solved then by decrease the z-score to increase the specify and decreasing false discovery rate. In addition, we proved that the accuracy increases by increasing the number of random drug sampling by factor 100 for our desired dataset. The other modification was related to genes without PPIs. Such genes are omitted from the PPI network, and in **Method and Discussion** we saw that ignoring such genes can lead to an increased false positive and false negative. And the last, Guney's study does not specify the p-value threshold used to classify valid drugs. In our approach, we considered drugs with a p-value ≤ 0.05 as valid. **Table 2** summarizes the modifications applied in Guney's method.

Table 1. Four statistical measures used for the comparison between **1k, 10k and 100k samples** revealed substantial discrepancies: Standard Deviation of Differences (SDD), Mean Difference, Range of Differences, and Coefficient of Variation (CV).

	z score	p-value	mean	SD
1k random sampling vs. 10k random sampling				
SDD	0.227	0.010	8.982	53.046
Mean Difference	0.002	-0.002	0.399	-0.474
Range of Differences	6.542	0.091	249.599	919.522
Coefficient of Variation (CV)	135.793	6.251	22.488	111.879
10k random sampling vs. 100k random sampling				
SDD	0.011	0.003	3.038	16.482
Mean Difference	0.000	0.000	-0.061	-1.977
Range of Differences	0.119	0.024	77.087	248.678
Coefficient of Variation (CV)	125.758	47.038	49.488	8.338
1k random sampling vs. 100k random sampling				
SDD	0.227	0.010	8.065	49.452
Mean Difference	0.002	-0.002	0.338	-2.451
Range of Differences	6.521	0.087	233.566	821.532

Coefficient of Variation (CV)	128.869	6.204	23.859	20.178

Table 2. Differences between the original and revised methods are listed

Method	Original method	Revised method
Drug-gene interaction score	Not applied	Applied
Genes with no PPI	Omitted	Considered
Random drug gene set sampling	1,000	100,000
Z-score	<= -0.15	<= -1.96
P-value	Not specified	<= 0.05
Applicable drug	Not checked	Checked
Side effects check	No	Yes
Drug ranking	No	Yes

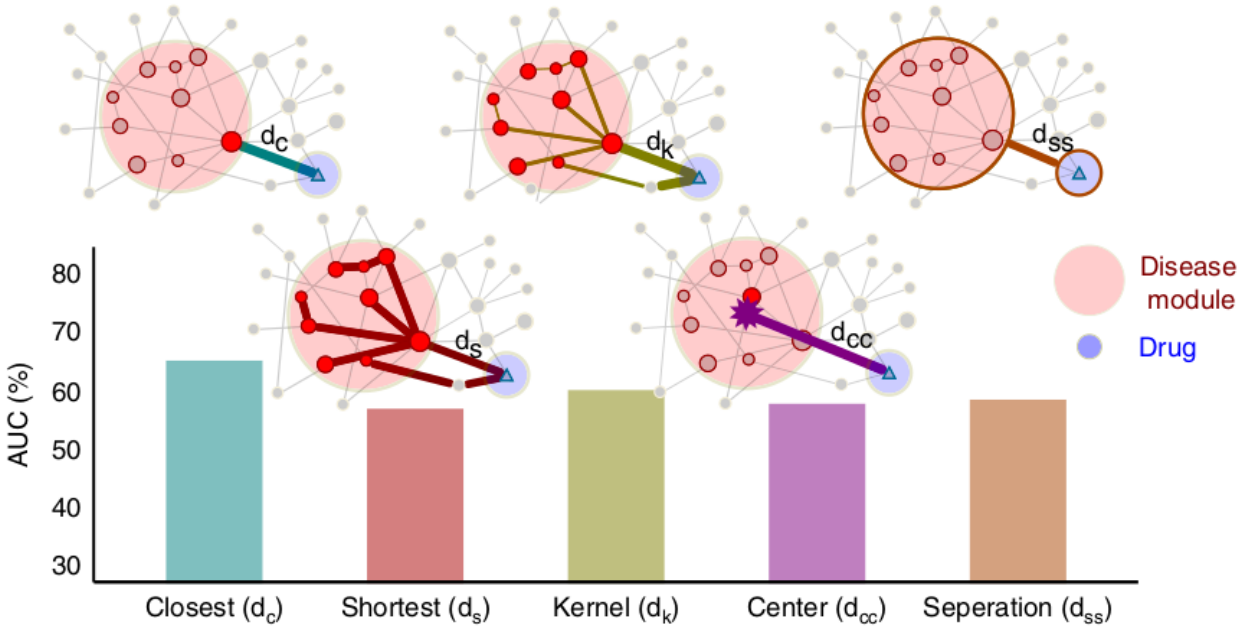

Figure 1. AUC is shown for relative proximity, z, calculated using five different distance measures. The closest measure, d_c, considers the shortest path length from each target to the closest disease protein, the shortest measure, d_s averages over all shortest path lengths to the disease proteins. See the text for the definition of the kernel (d_k), center (d_{cc}) and separation (d_{ss}) measures.

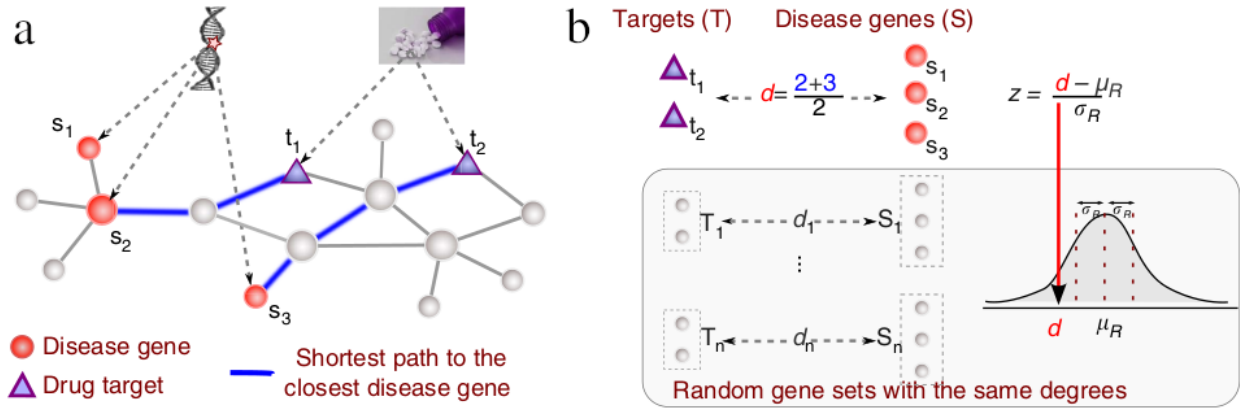

Figure 2. Illustration of the closest distance (d_c) of a drug T with targets t_1 and t_2 to the proteins s_1, s_2 and s_3 associated with disease S. To measure the relative proximity (zc), we compare the distance dc between T and S to a reference distribution of distances observed if the drug targets and disease proteins are randomly chosen from the interactome. The obtained proximity zc quantifies whether a particular d_c is smaller than expected by chance. To account for the heterogeneous degree distribution of the interactome and differences in the number of drug targets and disease proteins, we preserve the number and degrees of the randomized targets and disease proteins.

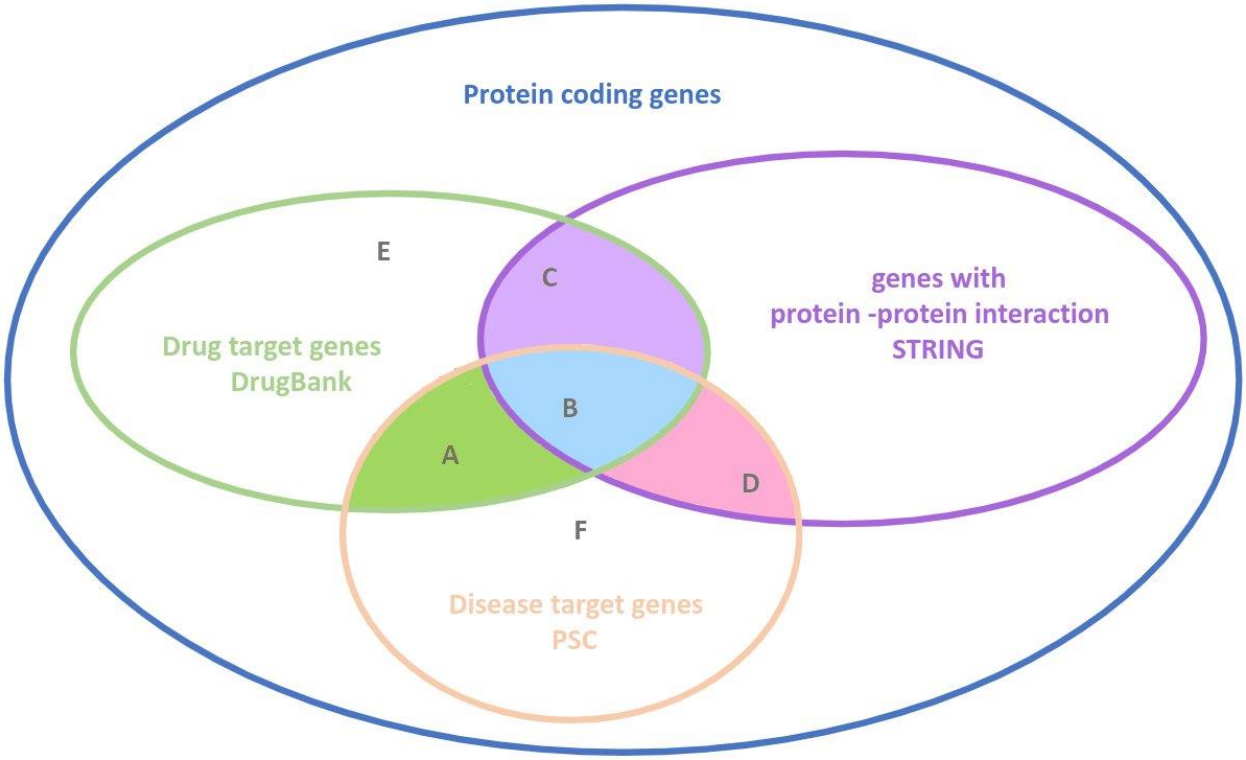

Figure 3. shows how drug target proteins and disease target proteins can be connected. Proteins in subset A and B are mutual protein between drug and disease target proteins. Drug proteins in subset C have interaction with some proteins but no connection with any disease protein. Disease proteins in subset D, have interaction with some proteins but no connection with any drug proteins. Subset E includes drug protein with any interaction, and Subset F includes disease protein with any interaction.

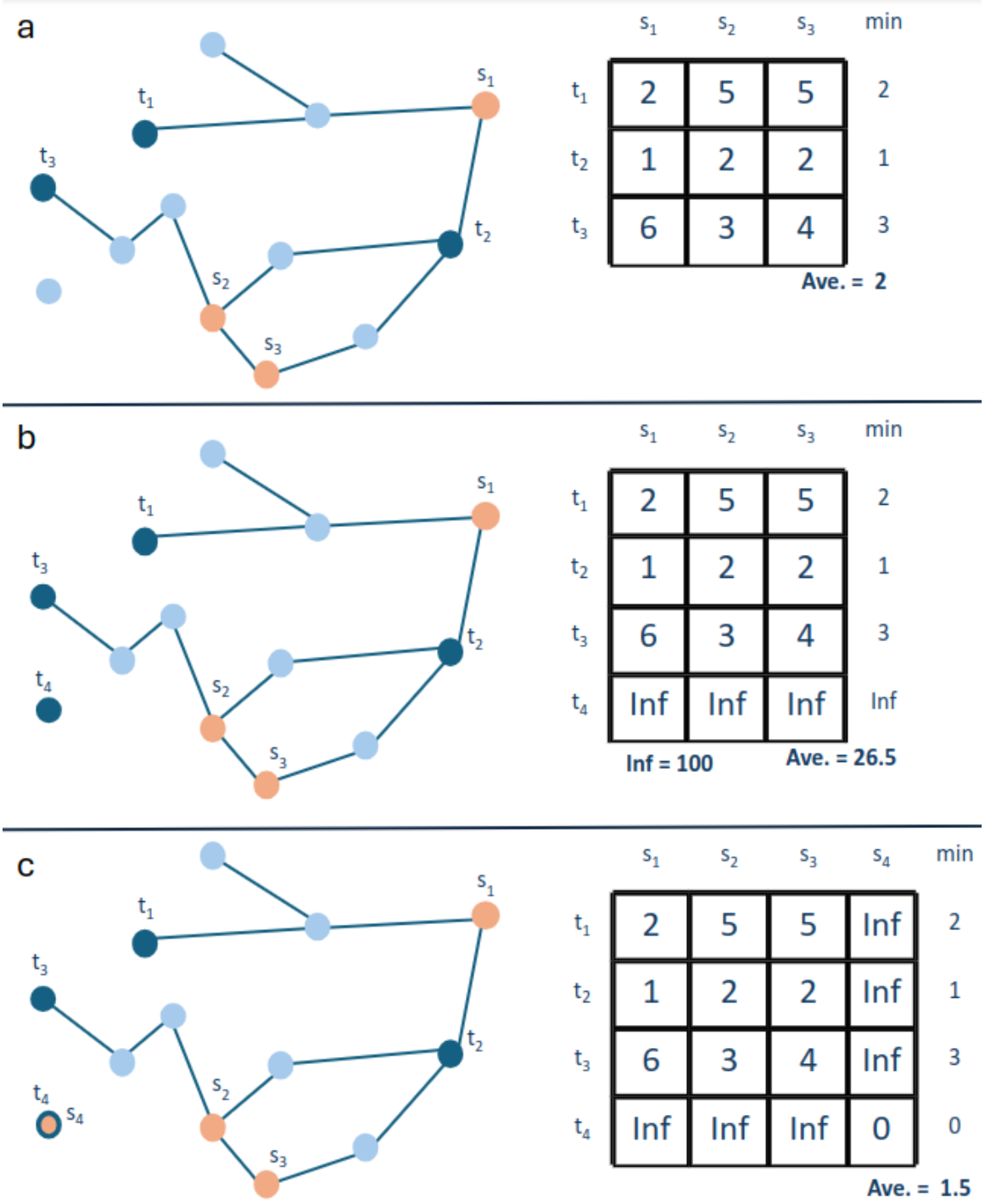

Figure 4. Indicates three examples a) the protein network with three drug target (shown by t_i) and three disease target proteins (shown by s_i) and all have protein interactions. b) shows a protein network that one drug target protein does not have protein interactions. c) represents a protein network that one drug target protein does not have protein interactions, but o it is a direct target of a disease protein.

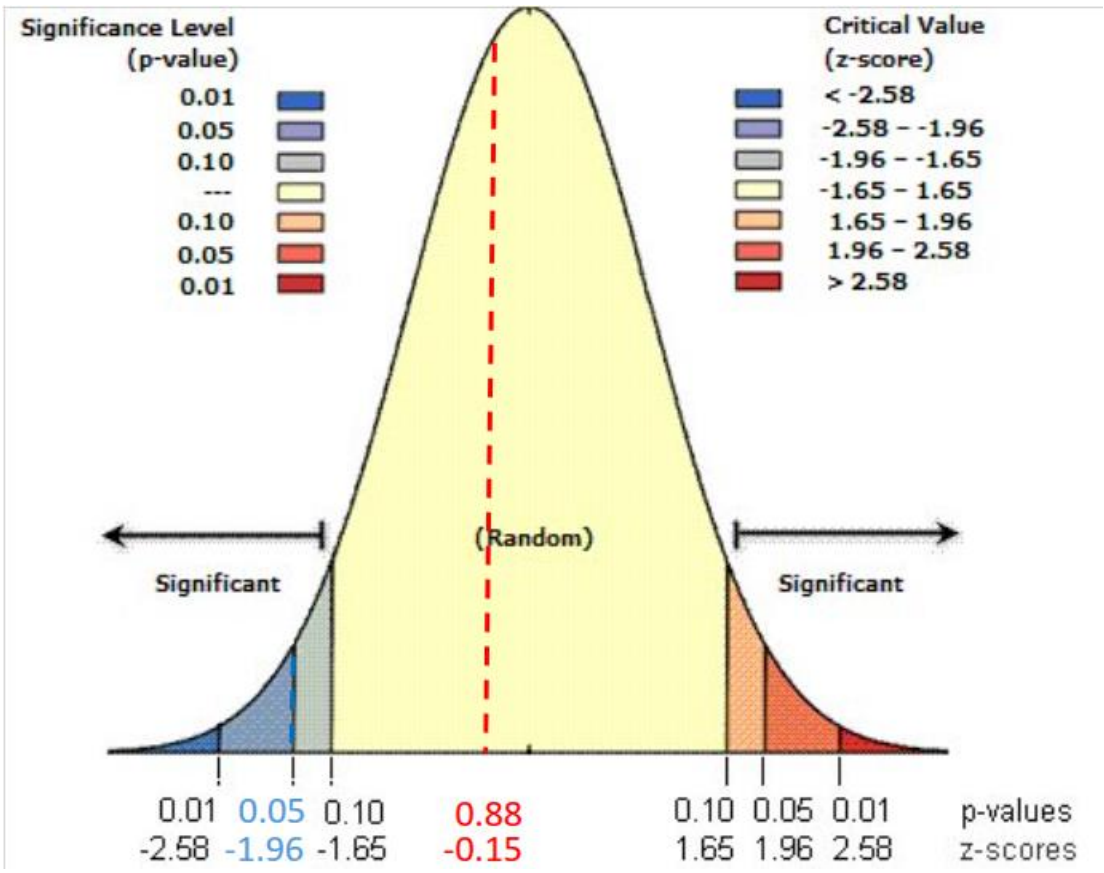

Figure 5. Z-score area under cave. The red dashed-line indicated the z-score(-0.15) used in Guney et al[cite] study. And the blue dashed-line shows the z-score is used in our method (-1.96).

```
#Pseudocode for calculating p-value and z-score for drug-disease gene sets

d = mean(shortest_distances(drug,disease)) # Calculate mean shortest distance for the observed drug-disease gene pairs
len(values) = random-drug-size # Total number of random drug selections
values = [] # List to store mean shortest distances for random drugs
for i = 1 to number_of_random_drugs:
    random_drug = getRandomDrug() # Select a random drug
    values[i] = mean(shortest_distances(random drug,disease)) # Calculate mean shortest distance for the i-th random
                                                              drug and the disease gene pairs

m = mean(values) # Mean of random drug distances
s = std(values) # Standard deviation of random drug distances
p-value = sum(values <= d)) / len(values) # Calculate the p-value based on random drug distances and expected mean
                                          shortest distance
z-score = (d - m) / s # Z-score calculation for the observed distance compared to the random distributions
```

Figure 6. Pseudocode for calculation p-value and z-score for desired drug and the disease gene set.

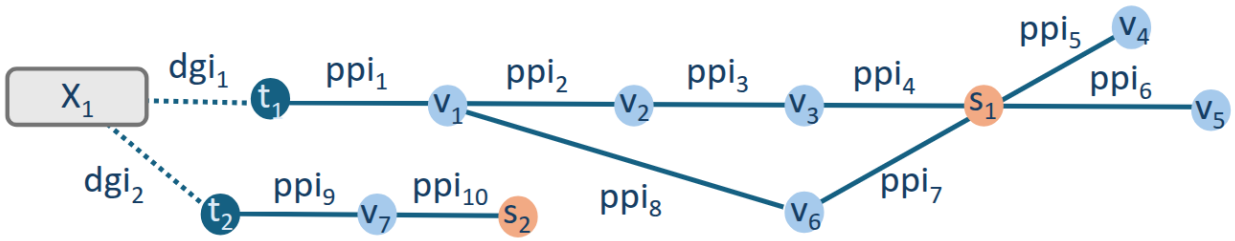

Figure 7. A toy PPI network with a sample drug X_1 and targeted genes t_1 and t_2. And a disease sample with target genes s_1 and s_2.

References