Supplementary Materials of "Prompt-based Generation of Natural Language Explanations of Synthetic Lethality for Cancer Drug Discovery"

Ke Zhang[1,2,3], Yimiao Feng[1,4], Jie Zheng[1,5,*]

[1] School of Information Science and Technology, ShanghaiTech University, Shanghai, China

[2]Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China

[3]University of Chinese Academy of Sciences, Beijing, China

[4]Lingang Laboratory, Shanghai, China

[5]Shanghai Engineering Research Center of Intelligent Vision and Imaging, ShanghaiTech University, Shanghai, China

{zhangke1, fengym, zhengjie}@shanghaitech.edu.cn

1 An example of data augmentation prompt

You are a helpful assistant that rephrases text and makes sentences smooth. I will give you a sample, please rephrase the partial sentence after the word "because" of the sample, then give me 10 rephrased answers. Each answer should include the exact noun phrases which I will give you, and each answer must start with "because". The complete sample is: *TP53 and CDK2 have a synthetic lethality relationship, because TP53 is a tumor suppressor that regulates cell cycle arrest, apoptosis and DNA repair, and CDK2 is a cyclin-dependent kinase that controls cell cycle progression and DNA replication. Therefore, inhibition of CDK2 in TP53-mutant cells results in synergistic cell death due to impaired DNA repair and increased DNA damage.* The phrases are *"DNA repair"*, *"DNA damage"*, *"cell cycle progression"*, *"cell death"*.

2 Human annotation pipeline

Algorithm 1: Human annotation

Input: GenepairsCollection Q_{nb}, AnswerCollection M_{nb}, CitationCollection R_{nb}
FactGenePairsCollection $Q_{fact} \leftarrow \{\}$
FactAnswerCollection $M_{fact} \leftarrow \{\}$
FactCitationCollection $R_{fact} \leftarrow \{\}$
FeatureCollection $F_{fact} \leftarrow \{\}$
HypotheticalAnswerCollection $M_{hypo} \leftarrow \{\}$
for $i \leftarrow 1$ **to** $len(M_{nb})$ **do**
    gene pair $(u, v) = Q_{nb}[j]$
    answer $\leftarrow M_{nb}$[i]
    citations $\leftarrow R_{nb}$[i]
    if AnnotatorReadandCheck(answer, citations) **then**
        features $\leftarrow$ FeatureAnnotation(answer)
        $Q_{fact} \leftarrow$ (u, v)
        $M_{fact} \leftarrow$ answer
        $R_{fact} \leftarrow$ citations
        $F_{fact} \leftarrow$ features
    else
        add answers to M_{hypo}
    end
    $pairs_{new} \leftarrow$ AnnotatorMiningNewpairs(citations)
    if $pairs_{new} \neq \varnothing$ **then**
        $answers_{new} \leftarrow$AnnotatorSummarization(citations)
        features $\leftarrow$ FeatureAnnotation(answers)
        $Q_{fact} \leftarrow pairs_{new}$
        M_{fact} $answers_{new}$
        $R_{fact} \leftarrow$ citations
        $F_{fact} \leftarrow$ features
    end
end

3 From a KG subgraph to a personalized KG prompt

According to the KG subgraph in Fig. 1, since RAD52 shares two functions with BRAC2's SL partner genes, we assume that RAD52 and BRAC2 also share the functions. Therefore, the KG prompt for BRAC2 and RAD52 is: *BRAC2* and *RAD52* may share common functions, including *DNA Damage Response*, *DNA repair*.

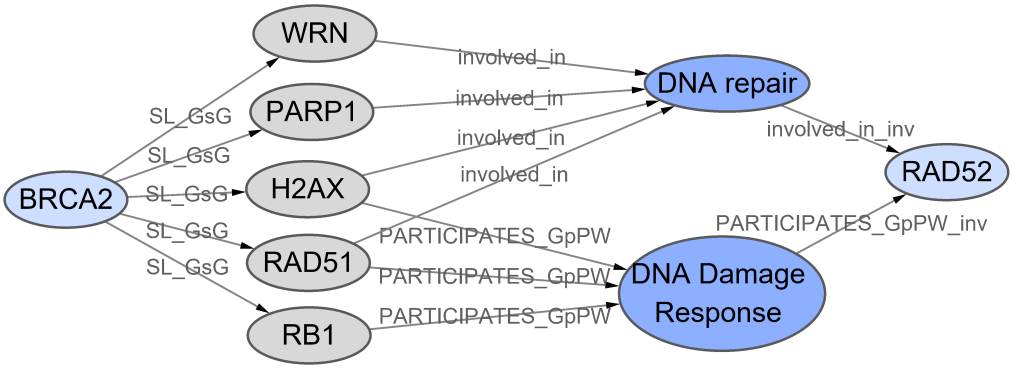

Figure 1: A KG subgraph for an SL gene pair (BRCA2, RAD52). Light blue nodes are the two target genes, grey nodes represent other genes that have SL relationships with BRAC2, and dark blue nodes are two gene functions.