Appendix II (A) 2. Prompt Generation Techniques

Phase I. Prompt for Pathway Description Summarization

Model: meta-llama/Meta-Llama-3.1-8B-Instruct.

Parameters: max_new_tokens=512, temperature=0.3, do_sample=False.

Instruction: Refine the provided biological pathway description by removing all alphanumeric IDs, URLs, and external references. If the description is brief or missing, create a factual summary based on recent literature, avoiding speculation. For lengthy descriptions, produce a concise summary that highlights the key functions and significance of the pathway. Provide only the refined text in one coherent paragraph of up to 300 words, omitting headers and details of deletions.

Content: "<Pathway Name>: <Pathway Description>"

Pathway Description Example: Sudden infant death syndrome (SIDS) susceptibility pathways: In this model, we provide an integrated view of Sudden Infant Death Syndrome (SIDS) at the level of implicated tissues, signaling networks and genetics. The purpose of this model is to serve as an overview of research in this field and recommend new candidates for more focused or genome wide analyses. SIDS is the sudden and unexpected death of an infant (less than 1 year of age), almost always during deep sleep, where no cause of death can be found by autopsy. Factors that mediate SIDS are likely to be both biological and behavioral, such as sleeping position, environment and stress during a critical phase of infant development (http://www.nichd.nih.gov/health/topics/Sudden_Infant_Death_Syndrome.cfm). While no clear diagnostic markers currently exist, several polymorphisms have been identified which are significantly over-represented in distinct SIDS ethnic population. The large majority of these polymorphisms exist in genes associated with neuronal signaling, cardiac contraction and inflammatory response. These and other lines of evidence suggest that SIDS has a strong autonomic nervous system component (PMID:12350301, PMID: 20124538). One of the neuronal nuclei most strongly implicated in SIDS has been the raphe nucleus of the brain stem. In this nuclei there are ultrastructural, cellular and molecular changes associated with SIDS relative to controls (PMID:19342987, PMID: 20124538). This region of the brain is responsible for the large majority of neuronal serotonin produced and is functionally important in the regulation of normal cardiopulmonary activity, sleep and thermoregulation (see associated references).

Genes associated with serotonin synthesis and receptivity have some of the strongest genetic association with SIDS. Principle among these genes the serotonin biosynthetic enzyme TPH2, the serotonin transporter SLC6A4 and the serotonin receptor HTR1A. SLC6A4 exhibits decreased expression in the raphe nucleus of the medulla oblongata and polymorphisms specifically associated with SIDS (PMID:19342987). In 75% of infants with SIDS, there is decreased HTR1A

expression relative to controls along with an increase in the number of raphe serotonin neurons (PMID:19342987). Over-expression of the mouse orthologue of the HTR1A gene in the juvenile mouse medulla produces an analogous phenotype to SIDS with death due to bradycardia and hypothermia (PMID:18599790). These genes as well as those involved in serotonin synthesis are predicted to be transcriptionally regulated by a common factor, FEV (human orthologue of PET-1). PET-1 knock-out results in up to a 90% loss of serotonin neurons (PMID:12546819), while polymorphisms in FEV are over-represented in African American infants with SIDS. In addition to FEV, other transcription factors implicated in the regulation of these genes (Putative transcriptional regulators (TRs)) and FEV are also listed (see associated references). In addition to serotonin, vasopressin signaling and its regulation by serotonin appear to be important in a common pathway of cardiopulmonary regulation (PMID:2058745). A protein that associates with vasopressin signaling, named pituitary adenylate cyclase-activating polypeptide (ADCYAP1), results in a SIDS like phenotype, characterized by a high increase in spontaneous neonatal death, exacerbated by hypothermia and hypoxia (PMID:14608012), when disrupted in mice. Protein for this gene is widely distributed throughout the central nervous system (CNS), including autonomic control centers (PMID:12389210). ADCYAP1 and HTR1A are both predicted to be transcriptionally regulated by REST promoter binding. Regulation of G-protein coupled signaling pathways is illustrated for these genes, however, it is not clear whether ADCYAP1 acts directly upon raphe serotonin neurons.

Another potentially important class of receptors in SIDS is nicotine. Receptors for nicotine are expressed in serotonin neurons of the raphe throughout development (PMID:18986852). Application of nicotine or cigarette smoke is sufficient to inhibit electrical activity of raphe serotonin neurons (PMID:17515803) and chronic nicotine infusion in rats decreases expression of SLC6A4 (PMID:18778441). Furthermore, nicotine exposure reduces both HTR1A and HTR2A immunoreactivity in several nuclei of the brainstem (PMID:17451658).

In addition to CNS abnormalities, several studies have identified a critical link between cardiac arrhythmia (long QT syndrome) and SIDS (PMID:18928334). A number of genetic association studies identified functionally modifying mutations in critical cardiac channels in as many as 10% of all SIDS cases (PMID:18928334). These mutations have been predicted to predispose infants for long QT syndrome and sudden death. The highest proportion of SIDS associated mutations (both inherited and sporadic) is found in the sodium channel gene SCN5A. Examination of putative transcriptional regulators for these genes, highlights a diverse set of factors as well as a relatively common one (SP1).

Finally, several miscellaneous mutations have been identified in genes associated with inflammatory response and thermoregulation. Infection is considered a significant risk factor for SIDS (PMID:19114412). For inflammatory associated genes, such as TNF alpha, interleukin 10 and complement component 4, many of these mutations are only significant in the presence of infection and SIDS. In addition to these mutations, cerebrospinal fluid levels of IL6 are increased in SIDS cases as well as IL6R levels in the arcuate nucleus of the brain, another major site of

serotonin synthesis (PMID:19396608). Genes such as ILR6 and ADCYAP1 are also associated with autoimmune disorders, thus SIDS may also be associated with autoinflammation of autonomic centers in the brain. Regulation of thermogenesis by brown adipose tissue has been proposed be an important component of SIDS, given that SIDS incidence is highest in the winter time and that animal models of SIDS demonstrate variation in body temperature. Interestingly, activation of raphe HTR1A decreases both shivering and peripheral vasoconstriction in piglets (18094064). Although a putative significant polymorphism was identified in the thermoregulator gene HSP60, this only occurred in one SIDS case. It is important to note that in the large majority of all these studies, sleeping position and smoking were among the most significant risk factors for SIDS.

In loving memory of Milo Salomonis (http://www.milosalomonis.org).

Proteins on this pathway have targeted assays available via the [https://assays.cancer.gov/available_assays?wp_id=WP706 CPTAC Assay Portal].

Generated Summary: Sudden infant death syndrome (SIDS) is the sudden and unexpected death of an infant under one year of age, often occurring during deep sleep, with no identifiable cause of death. Research suggests that SIDS is influenced by both biological and behavioral factors, including sleeping position, environment, and stress during a critical phase of infant development. Genetic studies have identified several polymorphisms associated with SIDS, primarily in genes related to neuronal signaling, cardiac contraction, and inflammatory response. The autonomic nervous system, particularly the raphe nucleus of the brain stem, is strongly implicated in SIDS, with ultrastructural, cellular, and molecular changes observed in this region. Genes associated with serotonin synthesis and receptivity, such as TPH2, SLC6A4, and HTR1A, have been linked to SIDS, with decreased expression of SLC6A4 and HTR1A observed in SIDS cases. Other factors, including vasopressin signaling, nicotine exposure, and cardiac arrhythmia, have also been implicated in SIDS. In addition, genetic mutations in genes associated with inflammatory response and thermoregulation have been identified, with infection and autoinflammation of autonomic centers in the brain potentially contributing to SIDS. Overall, SIDS is a complex condition influenced by multiple genetic and environmental factors, highlighting the need for further research to understand its underlying mechanisms and develop effective prevention strategies.

Phase II. Prompt for Embedding Generation

Type 1. Gene Symbol

Description: The prompt lists up to 100 gene symbols, sorted by their RP_SCORE in descending order, focusing on the most influential genes while excluding the less influential ones.

Example: 'EP300 CREBBP JUN HDAC1 SP1 PRKACA CTNNB1 CREB1 AR VEGFA IL6 TNF NR3C1 NFKB1 IL1B CEBPB HIF1A NGF GAPDH YWHAZ BDNF YWHAE HDAC9 TH PRKACB ESR2 YWHAH CASP3 IL10 EGR1 YWHAQ YWHAG MEF2C SST PPARGC1A HSPD1 PRKAR2A SP3 PRKAR2B RORA ADCYAP1 GRIN1 TPH1 PRKAR1A TCF3 HTR1A YWHAB CREM YBX1 TAC1 THRB IL1A NFKB2 VAMP2 POU5F1 SOX2 CPT1A ASCL1 IL13 MYB SPTBN1 GATA3 HES1 POU2F2 MECP2 PRKAR1B SNTA1 KCNH2 PHOX2A NANOG RET NFYA CHAT RYR2 IL1RN SCN5A CHRNA4 DDC HADHA RUNX3 GJA1 PBX1 MAOA CHRM2 GATA2 KCNQ1 MAP2 PHOX2B IL6R HTR2A ACADM AVP TACR1 SLC6A4 GCK NKX2-2 TP73 HTR3A HADHB MAZ'

Instruction for llm2vec models: "Given a list of gene symbols, encode the collective biological significance and pathway associations of these genes:"

Note: Along with the main prompt, the instruct models (Mistral-7B and Llama3-8B) require ‘instruction’ about the task to be performed. This instruction and prompt pair is fed to the model to extract the embeddings.

Type 2. Gene Description

Description: Similar to Type 1, this includes both Gene Symbols and their Descriptions.

Example: 'EP300 (E1A binding protein p300), CREBBP (CREB binding protein), JUN (Jun proto-oncogene, AP-1 transcription factor subunit), HDAC1 (histone deacetylase 1), SP1 (Sp1 transcription factor), PRKACA (protein kinase cAMP-activated catalytic subunit alpha), CTNNB1 (catenin beta 1), CREB1 (cAMP responsive element binding protein 1), AR (androgen receptor), VEGFA (vascular endothelial growth factor A), IL6 (interleukin 6), TNF (tumor necrosis factor), NR3C1 (nuclear receptor subfamily 3 group C member 1), ..., GCK (glucokinase), NKX2-2 (NK2 homeobox 2), TP73 (tumor protein p73), HTR3A (5-hydroxytryptamine receptor 3A), HADHB (hydroxyacyl-CoA dehydrogenase trifunctional multienzyme complex subunit beta), MAZ (MYC associated zinc finger protein)'

Instruction for llm2vec models: "Given a list of gene symbols and its description, encode the collective biological significance and pathway associations of these genes:"

Type 3. Pathway Name

Description: The prompt simply consists of the pathway name.

Example: Sudden infant death syndrome (SIDS) susceptibility pathways.

Instruction for llm2vec models: "Given a pathway name, encode the biological significance and functional associations of this pathway:"

Type 4. Pathway Description Summary

Description: The prompt consists of the summary generated from *I. Prompt for Pathway Description Summarization*.

Example: Sudden infant death syndrome (SIDS) is the sudden and unexpected death of an infant under one year of age, often occurring during deep sleep, with no identifiable cause of death. Research suggests that SIDS is influenced by both biological and behavioral factors, including sleeping position, environment, and stress during a critical phase of infant development. Genetic studies have identified several polymorphisms associated with SIDS, primarily in genes related to neuronal signaling, cardiac contraction, and inflammatory response. The autonomic nervous system, particularly the raphe nucleus of the brain stem, is strongly implicated in SIDS, with ultrastructural, cellular, and molecular changes observed in this region. Genes associated with serotonin synthesis and receptivity, such as TPH2, SLC6A4, and HTR1A, have been linked to SIDS, with decreased expression of SLC6A4 and HTR1A observed in SIDS cases. Other factors, including vasopressin signaling, nicotine exposure, and cardiac arrhythmia, have also been implicated in SIDS. In addition, genetic mutations in genes associated with inflammatory response and thermoregulation have been identified, with infection and autoinflammation of autonomic centers in the brain potentially contributing to SIDS. Overall, SIDS is a complex condition influenced by multiple genetic and environmental factors, highlighting the need for further research to understand its underlying mechanisms and develop effective prevention strategies.

Instruction for llm2vec models: "Given a pathway name and its description, encode the biological significance and functional associations of this pathway:"