information

Article

A Tissue-Specific and Toxicology-Focused Knowledge Graph

Ignacio J. Tripodi *, Lena Schmidt, Brian E. Howard, Deepak Mav and Ruchir Shah

Sciome, LLC, Research Triangle Park, Durham, NC 27709, USA
* Correspondence: ignacio.tripodi@sciome.com

Abstract: Molecular biology-focused knowledge graphs (KGs) are directed graphs that integrate information from heterogeneous sources of biological and biomedical data, such as ontologies and public databases. They provide a holistic view of biology, chemistry, and disease, allowing users to draw non-obvious connections between concepts through shared associations. While these massive graphs are constructed using carefully curated ontologies and annotations from public databases, much of the information relating the concepts is context specific. Two important variables that determine the applicability of a given ontology annotation are the species and (especially) the tissue type in which it takes place. Using a data-driven approach and the results from thousands of high-quality gene expression samples, we have constructed tissue-specific KGs (using liver, kidney, and heart as examples) that empirically validate the annotations provided by ontology curators. The resulting human-centered KGs are designed for toxicology applications but are generalizable to other areas of human biology, addressing the issue of tissue specificity that often limits the applicability of other large KGs. These knowledge graphs can serve as valuable tools for generating transparent explanations of experimental results in the form of mechanistic hypotheses that are highly relevant to the studied tissue. Because the data-driven relations are derived from a large collection of human in vitro data, these KGs are particularly well suited for in vitro toxicology applications.

Keywords: knowledge graphs; semantic web; knowledge mining; ontologies; semantic knowledge curation

Citation: Tripodi, I.J.; Schmidt, L.; Howard, B.E.; Mav, D.; Shah, R. A Tissue-Specific and Toxicology-Focused Knowledge Graph. *Information* **2023**, *14*, 91. https://doi.org/10.3390/info14020091

Academic Editors: Pierpaolo Basile and Annalina Caputo

Received: 19 October 2022
Revised: 20 January 2023
Accepted: 26 January 2023
Published: 3 February 2023

1. Introduction

Over the past several decades, great effort has been put into the careful curation of open biomedical ontologies and pathway databases as well as manual annotation of the relevant physical entities denoted by these concepts (e.g., proteins, genes, or chemicals). Ontologies provide a set of formal, hierarchical descriptions of abstract concepts within a highly specific scope, connected by precise relation types. Axioms are commonly defined them as "triples" (subject–predicate–object, i.e., a subject related to an object via a predicate), which helps data-mining and artificial-intelligence tools to exploit the structured knowledge computationally. Several popular ontologies and pathway databases, such as the Gene Ontology (GO) [1,2] and Reactome [3], are used quite commonly and have become staple downstream analysis endpoints for many critical tasks in 'omics, such as the functional characterization of sets of differentially expressed or mutationally altered genes. Despite the rich content of these individual resources, a natural limitation is that they provide a view of just one aspect of biology at a time. For example, we can find in ontology annotations that several genes (Entrez) participate in a biological process (GO); however, in a knowledge graph (KG), we can also see that some of their protein products (Uniprot [4]) molecularly interact (StringDB [5]) with one or more other proteins that participate in the same biochemical reaction (Reactome), which is a key component of certain disease-related (MONDO [6]) pathways. This kind of rich, heterogeneous information can only be derived from a knowledge aggregation data structure, such as a KG [7].

In order to break down these data silos, multiple ontologies and databases can often be combined into a comprehensive KG to provide a more holistic view of biology and

biochemistry. A semantic KG is a data model that structures knowledge as a directed graph for computational use, taking into account the specific relation types used to link abstract concepts and physical entities. Abstract concepts and physical entities are both defined as nodes in such a graph, and the relations between them are defined as directed edges connecting a pair of nodes (e.g., "BCL2L11 protein"—"participates in"–> "Translocation of BIM to mitochondria"). Several large KGs with diverse levels of granularity have been created for general biomedical applications [8], including KaBOB [9], Hetionet [10], and the example graph generated by the comprehensive KG creation framework, PheKnowLator [11]. A human-specific tissue-specific KG for toxicology applications, however, has so far been lacking. Furthermore, the annotations linking genes and proteins among themselves, as well as to different ontology concepts denoting biological processes, molecular functions, cellular compartments, pathways, phenotypes, or diseases can be very context specific. In particular, these types of relations may only be relevant for a specific tissue type.

To take a first step toward mitigating these shortcomings, we have produced a human-centered, toxicology-focused knowledge graph that can be tailored to different tissue types based on data-driven inference. This has been achieved by using a large curated collection of high-quality gene expression assays from a variety of tissues. For example, by using 3328 unperturbed liver samples to calculate the correlation of gene expression levels and single-sample concept enrichment scores (e.g., GO, Reactome, Human Phenotype Ontology (HPO) [12]), we can discard gene–>concept edges below a strong correlation threshold, under the assumption that there is no empirical evidence demonstrating that the gene really relates to the concept in liver tissue. Conversely, if the empirical evidence indicates consistent, strong tissue-specific correlation between the expression patterns of pair of genes, we can add a gene<–>gene edge to the KG to capture this relationship. The fact that the edges in the KG are empirically derived from in vitro data also positions these tissue-specific graphs uniquely well to analyze results from similar in vitro datasets, though they can also be employed for other types of data. We have so far produced a general human KG featuring data-driven edges, as well as tissue-specific KGs for liver, kidney and heart tissue using this method. The approach that we describe in the following sections can be applied to other tissues and organisms, provided that adequate experimental data are available.

2. Materials and Methods

2.1. KG Construction Process

Each KG was constructed from a combination of open biomedical ontologies, public databases, and data-driven relations produced from a large collection of in-house curated experimental datasets. The complete list of resources used can be found in Table 1.

For every resource incorporated into the KG, we pay careful attention to the most specific predicate type available, generally from the Relation Ontology (RO) [13]. The RO provides a formal framework to link ontology concepts, including both general types of relations and biomedical-specific relations. If the RO predicate class linking two concepts is provided by the ontology or annotation files, we use it directly as the edge type. Otherwise, we curate the most specific RO class for each new edge type we create. For example, for a new directed edge derived from the Toxin and Toxin-Target Database (T3DB) [14] connecting a chemical to a protein, we use the "regulates activity of" (RO:0011002) relation. Whenever the RO class allows it, an inverse edge is also added to traverse the pair of nodes in the opposite direction. We have relied on the RO class definitions to determine when an inverse relation can be added for each new edge. For example, when linking two subsequent biochemical reactions in a pathway using the "causally upstream of" relation (RO:0002411), we can also add a directed edge in the opposite direction of type "causally downstream of" (RO:0002404). The objective for this careful edge curation is twofold: First, to produce the most precise explanation when traversing the graph in search for evidence. Second, to make the KG ready to use with a semantic reasoner in the future, to infer additional edges via state-of-the-art deductive and inductive methods. The entire process described below is programmed as an automated pipeline.

Table 1. Public resources utilized for the creation of the KG. The node counts reflect the subset of concepts from each resource that were included in the KG.

Resource	Type of Knowledge Gathered
Gene Ontology (GO) (47,101 nodes)	Biological processes, molecular functions, and cellular compartments. Protein to GO concept relations. *Examples: fatty-acyl-CoA binding (GO:0000062), oxidoreductase activity, acting on metal ions (GO:0016722), endolysosome membrane (GO:0036020).*
Human Phenotype Ontology (HPO) (18,619 nodes)	Abnormal phenotypes. Gene to phenotype relations. Phenotype to disease relations. *Examples: Abnormal thrombocyte morphology (HP:0001872), Intrahepatic biliary atresia (HP:0005248).*
MONDO disease ontology (MONDO) [6] (22,398 nodes)	Diseases. Disease to phenotype relations. *Examples: Ullrich congenital muscular dystrophy (MONDO:0000355), pulmonary sarcoidosis (MONDO:0001708).*
Monarch Initiative [15]	Gene to disease relations. *Example: GTF2H5 (Entrez 404672) –contributes to condition (RO:0003304)–> trichothiodystrophy (MONDO:0018053).*
ClinVar [16]	Gene to disease relations. Gene to phenotype relations. *Example: AASS (Entrez 10157) –causes or contributes to condition (RO:0003302)–> Hyperlysinemia (HP:0002161).*
Chemicals of Biological Interest (ChEBI) [17] (168,563 nodes)	Chemicals, chemical groups and roles. *Examples: 1,2-dichloropropane (CHEBI:142468), phase-transfer catalyst (CHEBI:63060).*
Protein Ontology (PRO) [18] (73,668 nodes)	Proteins and protein families. *Example: nuclear factor NF-kappa-B p50 subunit (PR:000001757).*
Cell Ontology (CL) [19] (2527 nodes)	Cell types and anatomical references. *Example: stellate pyramidal neuron (CL:4023093).*
UniProt [4] (21,485 corresponding PRO nodes, 19,494 gene nodes)	Proteins and their corresponding gene templates. The human instance of the protein in PRO is used as identifier. *Example: CYP2E1 protein (PR:P05181).*
Reactome Pathway Database (28,898 nodes)	Biological pathways, and hierarchical relations between them. Biochemical reactions, and their relation to pathways. Protein complex relations to reactions and pathways. Protein and chemical participation in protein complexes. Gene relations to biological pathways. *Examples: B4GALT6 homodimer [Golgi membrane] (R-HSA-1015817), MAPK3, (MAPK1) phosphorylates GRB2-1:SOS1:p-Y427-SHC1 (R-HSA-109822), Activation of BIM and translocation to mitochondria (R-HSA-111446).*
StringDB [5]	Relations between proteins based on molecular interactions. We are only using those relations based on experimentally-validated physical interactions. The reported experimental score was used for the edge weight. *Example: NUD4B (PR:A0A024RBG1) –molecularly interacts with (RO:0002436)–> HDAC4 (PR:P56524).*
AOPwiki [20]	Relations between annotated AOP concepts from various ontologies. *Example: reactive oxygen species biosynthetic process (GO:1903409) –SCIOME:has_downstream_key_event (custom relation)–> oxidative stress (MP:0003674).*

Table 1. *Cont.*

Resource	Type of Knowledge Gathered
Toxin and Toxin-Target Database (T3DB) [14]	Relations between chemicals considered toxins and their target proteins. *Example: lead atom (CHEBI:25016) –regulates the activity of (RO:0011002)–> ATNG (PR:P54710).*
Relation Ontology (RO) [13] (41 relation types)	Formal description of relations between concepts and entities in the KG. *Examples: molecularly interacts with (RO:0002436), causes or contributes to condition (RO:0003302).*

For genes, we used the NCBI Entrez ID as the unique identifier, and for proteins, the PRO ID (the human instance represented by its UniProt ID, such as *PR:P37173* for protein *P37173*). Only human genes and proteins were used during the construction of these KGs.

In general terms, to construct the KGs, we started from the same steps. First we incorporated the nodes and edges from each ontology, which resulted in several large graphs with a few edges connecting each other. We then incorporated nodes and edges from public databases (i.e., genes, proteins, complexes, and pathways) and ontology "annotations" (lists of related genes and proteins to ontology concepts provided by ontology curators), which added many nodes and edges linking the different ontology concepts to common entities. Sometimes the same concept existed in multiple sources, so we simplified the KG by collapsing duplicate nodes into one, combining all their neighbors. Starting from this base KG, we used a data-driven approach to add edges between pairs of genes and remove gene edges to concepts that are not validated empirically. This last step was performed either with a large collection of samples from multiple tissue types to create a general KG, or samples from a specific tissue type to create tissue-specific KGs. As a general rule, any node incorporated into the KG was mapped via a standard, unique identifier (specific ontology ID, Uniprot ID, Entrez ID, etc.), and in the case of proteins and genes, only the human instance was preserved (ignoring non-human homologs and their annotations). The detailed steps followed to create the KG, listed by the overall types of connected nodes produced, were as follows:

1. **Add nodes and edges sourced from individual biomedical ontologies.** Nodes and edges are, in general, provided as triples, as described previously. This step may also include referenced nodes from external ontologies, such as Uberon Anatomy Ontology (UBERON), Phenotype And Trait Ontology (PATO), Cell Line Ontology (CLO), mammalian phenotype ontology (MP), etc.
    (a) $GO \leftrightarrow GO$: Add GO triples from the ontology Open Biological and Biomedical Ontology (OBO) definition. *Example: actin cortical patch assembly –is_a–> cellular component assembly (GO:0000147 –rdf-schema#subClassOf–> GO:0022607).*
    (b) $HPO \leftrightarrow HPO$: Add HPO triples from the ontology OBO definition. *Example: Hyperserinemia –is_a–> Abnormal circulating serine concentration (HP:0500138 –rdf-schema#subClassOf–> HP:0012278).*
    (c) $MONDO \leftrightarrow MONDO$: Add MONDO disease ontology triples from the ontology OBO definition. *Example: reticulate pigment disorder –is_a–> genetic skin disease (MONDO:0000118 –rdf-schema#subClassOf–> MONDO:0024255).*
    (d) $ChEBI \leftrightarrow ChEBI$: Add Chemicals of Biological Interest ontology (ChEBI) triples from the ontology OBO definition. *Examples: chloride –is_a–> halide anion (CHEBI:17996 –rdf-schema#subClassOf–> CHEBI:16042); chloride –is conjugate base of–> hydrogen chloride (CHEBI:17996 –chebi#is_conjugate_base_of–> CHEBI:17883).*
    (e) $PRO \leftrightarrow PRO$: Add PRO triples from the ontology OBO definition. *Example: LPS:GPI-anchored CD14 complex –has component–> lipopolysaccharide (PR:000025493 –RO:0002180–> CHEBI:16412).*

(f) $CL \leftrightarrow CL$: Add Cell Ontology (CL) triples from the ontology OBO definition. *Example: peridermal cell –is_a–> squamous epithelial cell (CL:0000078 –rdf-schema#subClassOf–> CL:0000076)*.

2. **Add nodes and edges from public databases.** In this step, we add nodes and edges derived from various biomedical databases. These are not ontologies defined in a semantic web format, but open databases that use a variety of data structures. They generally do not refer to abstract concepts like molecular functions or diseases, but rather to concrete entities, such as genes, proteins, chemicals, etc. Any nodes added to the KG were based on strict unique identifier rules. All proteins were identified by their human Uniprot IDs, ignoring the non-human homologs and their annotations. All human genes were identified by their Entrez IDs, and chemicals to their ChEBI IDs.

   (a) $gene \leftrightarrow protein$: Add gene-to-protein and protein-to-gene edges from UniProt, to define which protein is which gene product. *Example: PADI6 –has_gene_product–> Protein-arginine deiminase type-6 (Entrez 353238 –RO:0002205–> PR:Q6TGC4)*.

   (b) $protein \leftrightarrow protein$: Add protein-to-protein interaction edges from StringDB, only based on experimental evidence. The scaled experimental evidence score is used as the edge weight. *Example: 26S proteasome complex subunit SEM1 –molecularly_interacts_with–> Proteasome subunit alpha type-6 (PR:P60896 –RO:0002436–> PR:P60900)*.

   (c) $gene \leftrightarrow MONDO, gene \leftrightarrow HPO$: Add gene-to-MONDO or HPO edges from ClinVar annotations to incorporate information of genes implicated in disease or phenotypes. *Examples: ZIC2 –causes_condition–> holoprosencephaly 5 (Entrez 7546 –RO:0003303–> MONDO:0012322), ZIC2 –causes_or_contributes_to_condition–> Bilateral cleft lip (Entrez 7546 –RO:0003302–> HP:0100336)* .

   (d) $protein \leftrightarrow complex$, $chemical \leftrightarrow complex$: Add protein-to-protein complex and chemical-to-protein complex edges from Reactome, to incorporate information about complex members. *Example: Complement factor H –molecularly_interacts_with–> CFH:Host cell surface [plasma membrane] (PR:P08603 –RO:0002436–> R-HSA-1006173), heparins –molecularly_interacts_with–> CFH:Host cell surface [plasma membrane] (CHEBI:24505 –RO:0002436–> R-HSA-1006173)*.

   (e) $protein \leftrightarrow Reactome$: Add protein complex-to-pathway edges from Reactome to show which complexes participate in which pathways. *Example: ISGF3 bound to ISRE promotor elements [nucleoplasm] –participates_in–> Interferon alpha/beta signaling (R-HSA-1015697 –RO:0000056–> R-HSA-909733)*.

   (f) $Reactome \leftrightarrow Reactome$: Add pathway to pathway hierarchical edges (causally upstream/downstream pathways) from Reactome. *Example: Translesion synthesis by Y family DNA polymerases bypasses lesions on DNA template –causally_upstream_of–> Termination of translesion DNA synthesis (R-HSA-110313 –RO:0002411–> R-HSA-5656169)*.

   (g) $Reactome \leftrightarrow Reactome$: Add edges to connect biochemical reactions that take part in pathways Reactome. *Example: Cables1 links CDK2 and WEE1 –member_of–> Factors involved in megakaryocyte development and platelet production (R-HSA-1013881 –RO:0002350–> R-HSA-983231)*.

   (h) $protein \leftrightarrow Reactome$: Add edges for proteins that participate in biochemical reactions Reactome. *Example: Complex III subunit 3 –participates_in–> Electron transfer from ubiquinol to cytochrome c of complex III (PR:P00156 –RO:0000056–> R-HSA-164651)* .

   (i) $chemical \leftrightarrow Reactome$: Add edges for chemicals that participate in biochemical reactions from Reactome. *Example: aldehydo-L-iduronic acid –participates_in–> IDUA hydrolyses the unsulfated alpha-L-iduronosidic link in DS (CHEBI:28481 –RO:0000056–> R-HSA-1793186)*.

(j) *gene* ↔ *Reactome*: Add edges connecting genes annotated as biological pathway participants from Reactome. *Example: FGF4 –participates_in–> FGFRL1 modulation of FGFR1 signaling (Entrez 2249 –RO:0000056–> R-HSA-5658623)*.

3. **Add edges from ontology annotations.** Many of the biomedical ontologies provide curated annotations for the ontology terms, for example, describing how genes and proteins relate to them. These annotations can also describe relations to concepts from a different ontology.
   (a) *protein* ↔ *GO*: Add protein–>GO edges from GO annotations, to relate proteins with biological processes, molecular functions and cellular compartments. *Example: CYB5 –enables–> cytochrome-c oxidase activity (PR:P00167 –RO:0002327–> GO:0004129)*.
   (b) *gene* ↔ *HPO*: Add gene–>HPO edges from HPO annotations, to relate genes with their associated phenotypes. *Example: STS –causes_or_contributes_to_condition–> Abnormal stomach morphology (Entrez 412 –RO:0003302–> HP:0002577)*.
   (c) *gene* ↔ *MONDO*: Add gene–>MONDO edges from Monarch annotations to incorporate information about genes known to cause or contribute to diseases. *Example: CTSF –causes_or_contributes_to_condition–> adult neuronal ceroid lipofuscinosis (Entrez 8722 –RO:0003302–> MONDO:0019260)*.
   (d) *MONDO* ↔ *HPO*: Add disease–>phenotype edges from MONDO and HPO annotations. Weight these edges based on the frequency at which a phenotype is manifested in a disease, using the values provided in the annotations. *Example: muscular dystrophy-dystroglycanopathy –has_phenotype–> Seizure (MONDO:0000171 –RO:0002200–> HP:0001250)*.
   (e) Add edges connecting nodes in the KG key event relations in AOPwiki, for which a related ontology concept has been annotated. The edge weight is based on the evidence code or quantitative understanding score (if given). *Example: hyperplasia –has_upstream_key_event–> cell proliferation (MONDO:0005043 –aop_ontology#has_upstream_key_event–> GO:0008283)*.
   (f) *chemical* ↔ *protein*: Add chemical to protein edges from the T3DB, to link known chemical stressors with their known dysregulated proteins. *Example: metixene –regulates_activity_of–> Muscarinic acetylcholine receptor M4 (CHEBI:51024 –RO:0011002–> PR:P08173)*.

4. **Simplify the KG by removing redundant nodes.** Collapse any group of nodes with an identical label into a single new node. All inbound or outbound edges from the collapsed nodes are added to the new one (Figure 1). This reduces unnecessary redundancy in the graph and helps avoid knowledge fragmentation. The priority given to node types to define which is the identifier that remains as the new node name is based on the following order: MONDO, Reactome, HPO, GO, PATO, UBERON, then any other node. Additionally, collapse the taxon-neutral and taxon-specific (human) protein nodes from PRO into one (keeping the human instance node), to avoid unnecessary nodes since this is an organism-specific KG. No proteins from any taxa other than human were incorporated in this KG.
5. **Add/Remove data-driven edges to create the general and tissue-specific KGs.** In this step, we remove any edges between genes and pathways or ontology concepts that are not strongly correlated among the experimental samples used in this empirical step. Additional edges are added between pairs of genes with strongly correlated expression across many experimental conditions (Figure 2).
   (a) *gene* ↔ *gene*: Add gene-to-gene edges from an extensive and carefully curated collection of high-quality control human gene expression samples. We first calculated Pearson correlation coefficients between each gene pair across. The lower and upper significance thresholds were derived from the distribution of correlation coefficient values. Specifically, lower threshold was defined as 25th quartile minus 3 times the inter-quartile range (IQR) and upper

threshold was defined as the 75th percentile minus 3 times the IQR. The new edge between two gene node was added if the corresponding coefficient was greater(less) than the upper(lower) significance threshold. To specify new edges in the graph indicating direct or inverse correlation in expression, respectively, with the correlation coefficient between the gene pair used as the edge weight.

(b) *gene* ↔ *concept*: Remove and adjust the weights of edges between gene nodes and nodes denoting GO, HPO, MONDO, or Reactome concepts according to a data-driven approach. We first performed Single-Sample Gene Set Enrichment Analysis (SSGSEA) [21] to derive enrichment scores (ES) for each concept using high-quality curated control gene expression samples spanning many tissue types. Next, we derived Pearson correlation coefficients between each gene's normalized expression value and the related concept's enrichment scores. The average of all pairwise correlation is used as the significance threshold. Any existing gene–>concept edge with correlation below this significance threshold is removed from the KG, and the correlation coefficient is used as the edge weight for those remaining gene–>concept edges.

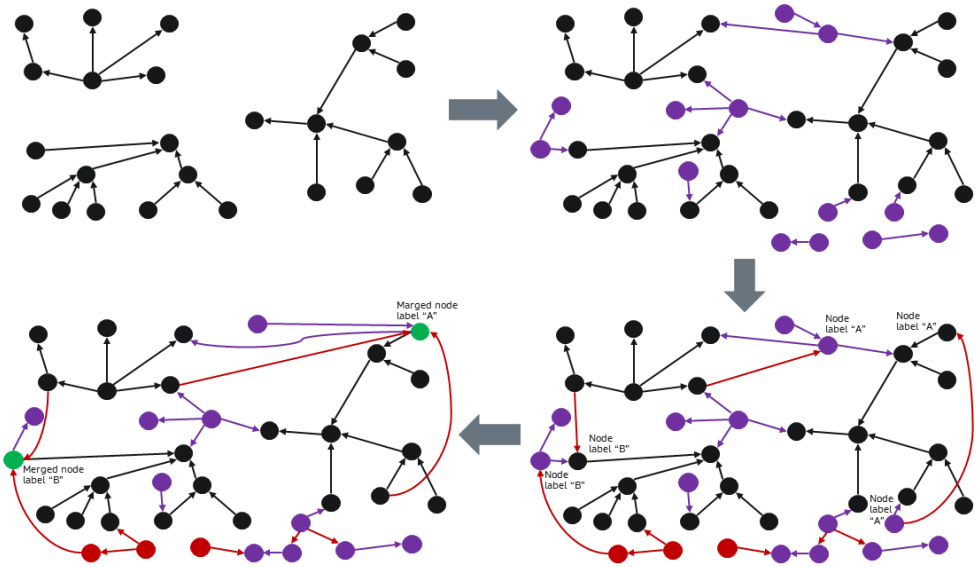

Figure 1. KG creation process. The different sources of knowledge are incorporated as nodes and edges, and those nodes with identical labels are collapsed into a single node, preserving the combination of edges. The data-driven edges are then added, linking genes among themselves or removing edges between genes and protein nodes and their annotations when there is not sufficient empirical evidence of strong interaction observed in vitro.

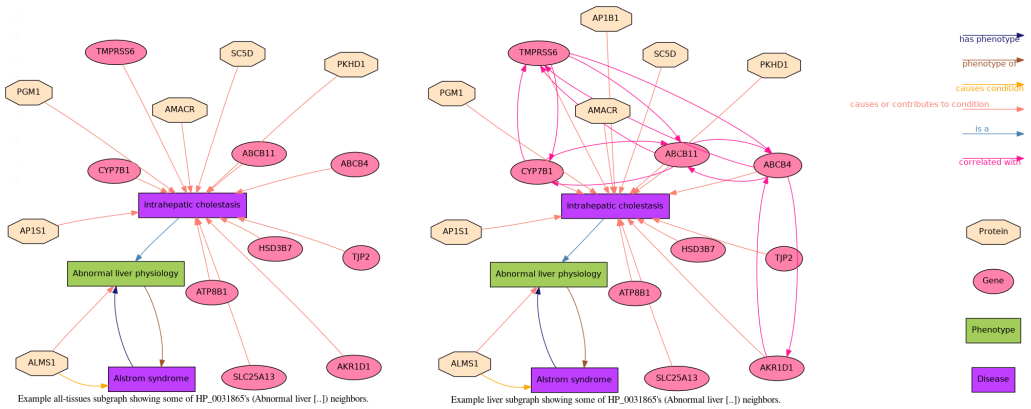

Figure 2. Example subgraph from the general (all tissues) KG compared to the same set of nodes in the liver KG. The handful of nodes displayed here illustrate the richness of relations among concepts and entities. The subgraph in the liver-specific KG includes several additional gene<->gene edges which are only derived empirically from liver samples, not when using all samples across tissues. It also features an additional protein edge connecting to one of the disease nodes (*AP1B1* to *intrahepatic cholestatsis*).

2.2. KG Statistics and Test Cases

We used this process to create a multi-tissue human KG from 10,000 control samples from a variety of tissues, to capture interactions generally reflected in a wide variety of in vitro experiments. In addition, we also created tissue-specific KGs by repeating the final step of our pipeline ("Step 5. Add/Remove data-driven edges to create the general and tissue-specific KGs") only utilizing data from that particular tissue type. As relevant examples to toxicology applications, we produced a liver-specific KG using 3328 gene expression in vitro samples from untreated, healthy liver tissue, as well as a kidney-specific KG from 843 kidney samples and a heart-specific KG from 711 samples. While we chose these organs as use cases for a tissue-specific KG due to their wide range of applications in toxicology, this step can be performed for any other tissue type with a sufficient number of samples.

To test an application of these tissue-specific graphs, we generated custom gene sets for concepts relevant to the tissue in question (e.g., node HP:0001395, "Hepatic fibrosis", to test the liver tissue KG). For each applicable concept, we tested two possible gene sets, one using the general (all tissues) KG relations and another one using the tissue-specific KG relations. Since the number of edges connecting genes or proteins to these concepts may vary among the different KGs, the resulting gene sets also differed and yielded different enrichment statistics. We compared the liver vs. general KG using the samples from an alcoholic hepatitis study [22] (GEO accession GSE28619), whereas a different study on nephrosclerosis [23] (GEO accession GSE20602) was used as a different example to compare the kidney vs. general KG.

3. Results

3.1. Resulting KGs

The resulting human-centered graph including the data-driven edges derived from all tissues contains 388,823 nodes and 4,270,374 directed edges. The edges between genes indicating either direct or inverse correlation, as well as the edges between genes (or proteins) and ontology concepts, were derived from a collection of 10,000 gene expression control (untreated) samples (across a wide variety of tissue types).

The number of edges in each tissue-specific KG will differ and depend on the number of samples available for that tissue, as well as the distribution of correlation coefficients between gene pairs and gene–>concept pairs. In general, as the number of samples used increases (thus making more correlation coefficients congregate closer to the distribution mean), the resulting graph will contain fewer data-driven edges with outlying correlation

weights. In the general KG derived from 10,000 samples, the universal interactions will be favored due to the strong tissue diversity. The liver-specific KG contains 5,931,485 directed edges, where its gene<->gene, gene–>concept or protein–>concept edges were derived from 3328 healthy, untreated liver samples. We also created a kidney-specific KG that contains 3,406,659 directed edges (derived from 843 kidney samples) and a heart-specific KG with 4,061,181 edges (derived from 711 heart samples).

3.2. Evaluation of Tissue-Specific KGs Using Edge Specificity

In the all-tissue KG, many of the gene–>concept edges that are available from public ontologies and databases are likely to correspond to relations that are tissue- and condition-specific. A goal of "Step 5: Add/Remove data-driven edges" is to use experimental gene expression data to create tissue-specific KGs by eliminating gene–>concept relations that are not supported in those tissues. To validate this approach, we hypothesized that when using gene expression data to prune the all-tissue KG, fewer gene–>concept edges should be lost for concept nodes explicitly relevant to the specific tissue of interest compared to the remainder of the nodes.

We first compiled lists of tissue-related concept nodes based on their labels. For example, after looking for concept nodes, including the (case-insensitive) strings "liver" or "hepat", we produced a list of 564 nodes that we then used to test the effect of liver specificity (see Supplemental Materials for the full list). The list included nodes such as *GO:0072575* (epithelial cell proliferation involved in liver morphogenesis), *HP:0006566* (Neonatal cholestatic liver disease), *MONDO:0003378* (liver leiomyosarcoma), *R-HSA-549129* (OCT1 transports organic cations into hepatic cells), etc. Similarly, the list of kidney-related nodes was constructed by searching for node labels containing the strings "kidney" or "renal" and manually removing false matches, resulting in 946 nodes. We used the strings "heart" and "cardi" to construct a list of 1461 heart-related nodes.

We can appreciate in Figures 3–5 how the data-driven method to prune edges for a specific tissue does indeed preferentially preserve many relations between gene or protein annotations and concepts related to the tissue of interest, while removing some of the gene–>concept edges incident to other concept nodes. These results are statistically significant; we confirm that the percentage of gene and protein neighbors lost to tissue-specific nodes is significantly lower ($p = 0.030$ on Welch's t-test for liver in Figure 3; $p = 1.53 \times 10^{-4}$ for kidney in Figure 4; and $p = 1.45 \times 10^{-9}$ for heart in Figure 5). In all cases, the gene or protein edges to nodes that are highly relevant to the specific KG tissue are largely preserved, while, in contrast, many edges are lost for the remaining nodes.

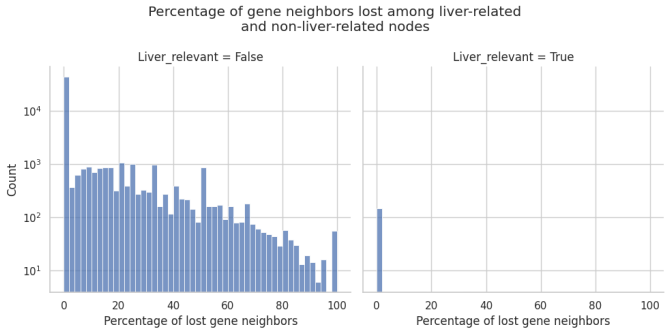

Figure 3. Distribution of the edges lost to gene and protein node neighbors for the liver-specific KG. Given a list of liver-related nodes (based on their label), we calculate the percentage of gene- or protein-adjacent node neighbors lost (**right panel**). When compared to the percentage of adjacent gene/protein neighbors lost for the remainder of (i.e., non-liver-related) nodes, we can see a significant increase in gene–>concept edges lost, many of them losing up to all gene or protein neighbors (**left panel**).

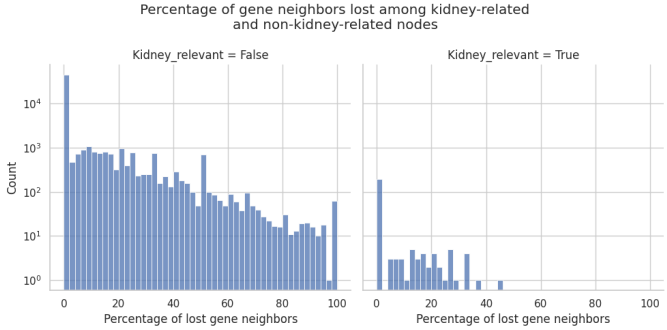

Figure 4. Distribution of the edges lost to gene and protein node neighbors for the kidney-specific KG. Given a list of kidney-related nodes (based on their label), we calculate the percentage of gene- or protein-adjacent node neighbors lost (**right panel**). When compared to the percentage of adjacent gene/protein neighbors lost for the remainder of (i.e., non-kidney-related) nodes, we can see a significant increase in gene–>concept edges lost, many of them losing up to all gene or protein neighbors (**left panel**).

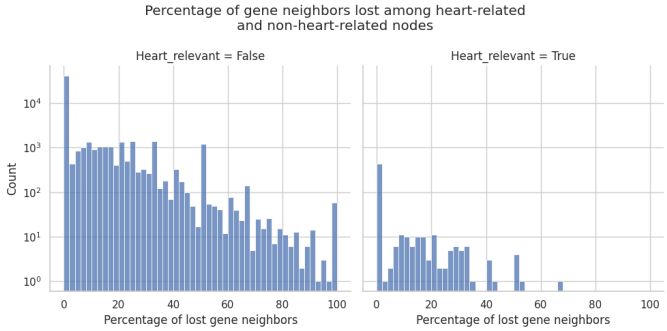

Figure 5. Distribution of the edges lost to gene and protein node neighbors for the heart-specific KG. Given a list of heart-related nodes (based on their label), we calculate the percentage of gene or protein adjacent node neighbors lost (**right panel**). When compared to the percentage of adjacent gene/protein neighbors lost for the remainder of (i.e., non-heart-related) nodes, we can see a significant increase in gene–>concept edges lost, many of them losing up to all gene or protein neighbors (**left panel**).

3.3. Evaluation of Tissue-Specific KGs Using Gene Set Enrichment

Next, we sought to test whether gene sets derived from tissue-specific knowledge graphs can improve our ability to identify concept enrichment from experimental data. As an illustrative test of a tissue-specific KG application, we calculated Gene Set Enrichment Analysis (GSEA) normalized enrichment scores (NESs) and statistical significance (p-value) of different concepts denoted by nodes in the KG, using their neighboring genes as gene sets. For each concept, we contrasted the NES and p-value obtained using the general KG with the equivalent values obtained using the tissue-specific KG. To calculate enrichment of these custom gene sets, we used a Python implementation [24] of the GSEA [25] algorithm. We used the t-test method to score any custom gene sets with a minimum of 3 genes, running 1000 phenotype-based permutations to assess statistical significance.

Using samples from a differential gene expression study that compared liver tissue affected by alcoholic hepatitis to healthy liver, we found that while the majority of related concepts' NES and p-value remained the same using either KG, four concepts that were deemed insignificant (p-value > 0.05) using gene sets derived from the general KG were significant and relevant when deriving the gene sets from the liver KG (Table 2): *liver disorder*, *response to ethanol*, *N-acylphosphatidylethanolamine metabolic process* (a biological process closely tied to alcohol intake [26]), and *chronic hepatic failure*. The only nodes that resulted in a slightly worse p-value are still significant and do not change the overall enrichment results.

Similarly, we compared NES and *p*-values of a study that compared nephrosclerosis kidney samples to healthy tissue and found two concepts that resulted in being insignificant using the general KG but were statistically significant using the kidney KG-derived gene sets (Table 3): *nephrosis* and *nephrotic syndrome*. Together, these two examples illustrate how tissue specificity of the gene to concept relations in our KGs can increase the power to detect concept enrichment using empirical data.

Table 2. Comparison of concept enrichment score between gene sets derived from the general KG vs. gene sets derive from the liver KG. The first four rows correspond to concepts that are highly relevant to the alcoholic hepatitis study and only become statistically significant when using the liver KG.

Concept ID	Concept Label	General NES	General p-val	Liver NES	Liver p-val
MONDO:0005154	**liver disorder**	**0.9318**	**0.52**	**1.4832**	**0.049**
GO:0070292	**N-acylphosphatidylethanolamine metabolic process**	**1.4047**	**0.11**	**1.6034**	**0.00593**
GO:0045471	**response to ethanol**	**1.5608**	**0.0509**	**1.6143**	**0.0154**
HP:0100626	**Chronic hepatic failure**	**1.5198**	**0.0594**	**1.5147**	**0.0370**
GO:0004022	alcohol dehydrogenase (NAD+) activity	1.6386	0.00204	1.6386	0.00204
GO:0004024	alcohol dehydrogenase activity, zinc-dependent	1.5774	0.012	1.5774	0.012
GO:0070291	N-acylethanolamine metabolic process	1.7220	0.00205	1.7220	0.00205
MONDO:0002520	hepatic porphyria	1.6078	0.0237	1.6078	0.0237
MONDO:0004721	liver neoplasm	1.6138	0.0174	1.6138	0.0174
MONDO:0007079	alcohol dependence	1.5304	0.00212	1.5304	0.00212
MONDO:0021698	alcohol-related disorders	1.6348	0	1.6348	0
R-HSA-71707	ethanol + NAD+ => acetaldehyde + NADH + H+	1.5281	0.0123	1.5281	0.0123
R-HSA-71384	Ethanol oxidation	1.6299	0.0156	1.6299	0.0156
GO:0006066	alcohol metabolic process	1.6005	0	1.4869	0.00375
MONDO:0019072	intrahepatic cholestasis	1.7858	0	1.7865	0.00699
GO:0006067	ethanol metabolic process	1.6299	0.0156	1.5795	0.0291

Table 3. Comparison of concept enrichment score between gene sets derived from the general KG vs. gene sets derive from the kidney KG. The first two rows correspond to concepts that are highly relevant to the nephrosclerosis study and only become statistically significant when using the kidney KG.

Concept ID	Concept Label	General NES	General p-val	Kidney NES	Kidney p-val
MONDO:0002331	**nephrosis**	**−0.7478**	**0.839**	**1.5524**	**0.00612**
MONDO:0005377	**nephrotic syndrome**	**−0.7478**	**0.839**	**1.5524**	**0.00612**
MONDO:0044765	steroid-resistant nephrotic syndrome	1.4808	0.0231	1.4740	0.015
GO:0072277	metanephric glomerular capillary formation	1.3856	0.0287	1.3856	0.0287
GO:0072557	IPAF inflammasome complex	1.5826	0	1.5826	0
GO:0072559	NLRP3 inflammasome complex	1.4586	0.0471	1.4586	0.0471
GO:0097169	AIM2 inflammasome complex	1.6029	0.00212	1.6029	0.00212
HP:0001685	Myocardial fibrosis	1.5911	0.011	1.5911	0.011
HP:0012593	Nephrotic range proteinuria	1.3798	0.0261	1.3798	0.0261
R-HSA-1234176	Oxygen-dependent proline hydroxylation of Hypoxia-inducible Factor Alpha	1.7026	0	1.7026	0
R-HSA-5678895	Defective CFTR causes cystic fibrosis	1.6973	0.00215	1.6973	0.00215
GO:0061702	inflammasome complex	1.5318	0.014	1.5043	0.0215

4. Discussion

Ontologies and pathway databases provide a wealth of knowledge; however, this knowledge is often very context dependent. Our data-driven approach allows us to focus on the more universal relations between genes (or the proteins they synthesize) and various types of concepts and pathways (from GO, Reactome, HPO, MONDO, etc.). It also allows

us to narrow the focus of the constructed KG to the tissue type under study, making it a uniquely capable enrichment tool for multi-omics downstream analysis. Deriving the tissue-specific relations from a large number of high-quality in vitro datasets positions, these graph uniquely well for downstream analysis of in vitro data when we seek enrichment of concepts and pathways.

A limitation of these KGs, like any others created under the "open-world" assumption, is that they are intended to reflect the axioms we currently know of. However, any missing edges do not necessarily imply that the corresponding relations are false; rather, we may simply lack the information required to make those assertions. Another limitation of the current approach is that when inferring data-driven edges, certain pathways may only be enriched by a subset of genes under very specific conditions. These rare conditions may be lost in the larger pool of samples for a given tissue, or it may simply be the case that no samples are available that are representative of those conditions.

The KGs we have produced can be expanded in different ways using other types of inferred edges in addition to the current data-driven ones. For example, the application of semantic reasoners for inductive learning can help us to detect generalizable rules that can help to infer additional facts or eliminate erroneous ones in the graph. Furthermore, an exercise equivalent to data-driven edge detection between genes can be conducted with other types of assays. For example, a collection of curated proteomics and chromatin accessibility samples could be used, along with tissue type, to infer context-dependent relations between proteins and, specifically, transcription factors.

Our unique approach to constructing semantic KGs tailored to specific tissue types, driven by empirical validation using human in vitro data, provides a resource ideally suited for analyzing new data from human in vitro experiments. It allows researchers to focus on context-specific downstream analysis of their 'omics experiments (as well as other in vitro techniques). By integrating biomedical knowledge across semantic domains, these KGs allow researchers to extract as much useful information as possible from human-relevant, time-efficient in vitro studies performed at large volume. In the particular case of toxicology, our combination of the expert knowledge incorporated into these graphs and data-driven inference from tens of thousands of assays in human tissue will drive the field of risk assessment forward by offering a holistic resource to be exploited when seeking the enrichment of adverse responses to chemical stressors.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info14020091/s1, (lists of tissue-related test KG nodes).

Author Contributions: Conceptualization, I.J.T.; software, I.J.T., L.S.; validation, I.J.T., L.S.; data curation, I.J.T.; writing—original draft preparation, I.J.T.; writing—review and editing, I.J.T., B.E.H., L.S., D.M., R.S.; visualization, I.J.T.; formal analysis, I.J.T., D.M.; investigation: I.J.T.; methodology: I.J.T., L.S., B.E.H., D.M.; supervision, B.E.H., R.S.; project administration, I.J.T. All authors have read and agreed to the published version of the manuscript.

Funding: None to report.

Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AOP	Adverse Outcome Pathway
ChEBI	Chemicals of Biological Interest Ontology
CL	Cell Ontology
CLO	Cell Line Ontology
GO	Gene Ontology

GSEA	Gene Set Enrichment Analysis
HPO	Human Phenotype Ontology
KG	Knowledge Graph
NES	Normalized Enrichment Score
OBO	Open Biological and Biomedical Ontology
PATO	Phenotype And Trait Ontology
PRO	Protein Ontology
RO	Relation Ontology
SSGSEA	Single-Sample Gene Set Enrichment Analysis
T3DB	Toxin and Toxin-Target Database
UBERON	Uberon Anatomy Ontology

References

1. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef] [PubMed]
2. Carbon, S.; Douglass, E.; Good, B.M.; Unni, D.R.; Harris, N.L.; Mungall, C.J.; Basu, S.; Chisholm, R.L.; Dodson, R.J.; Hartline, E.; et al. The Gene Ontology Resource: Enriching a gold mine. *Nucleic Acids Res.* **2020**, *49*, D325–D334. [CrossRef]
3. Gillespie, M.; Jassal, B.; Stephan, R.; Milacic, M.; Rothfels, K.; Senff-Ribeiro, A.; Griss, J.; Sevilla, C.; Matthews, L.; Gong, C.; et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **2022**, *50*, D687–D692. [CrossRef]
4. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515. [CrossRef]
5. Szklarczyk, D.; Gable, A.L.; Nastou, K.C.; Lyon, D.; Kirsch, R.; Pyysalo, S.; Doncheva, N.T.; Legeay, M.; Fang, T.; Bork, P.; et al. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **2021**, *49*, D605–D612. [CrossRef]
6. Vasilevsky, N.A.; Matentzoglu, N.A.; Toro, S.; Flack, J.E.; Hegde, H.; Unni, D.R.; Alyea, G.F.; Amberger, J.S.; Babb, L.; Balhoff, J.P.; et al. Mondo: Unifying Diseases for the World, by the World, 2022. Available online: http://purl.obolibrary.org/obo/mondo.obo (accessed on 23 May 2022).
7. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Yu, P.S. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 494–514. [CrossRef]
8. Li, M.M.; Huang, K.; Zitnik, M. Graph representation learning in biomedicine and healthcare. *Nat. Biomed. Eng.* **2022**, *6*, 1353–1369. [CrossRef]
9. Livingston, K.M.; Bada, M.; Baumgartner, W.A.; Hunter, L.E. KaBOB: Ontology-based semantic integration of biomedical databases. *BMC Bioinform.* **2015**, *16*, 126. [CrossRef]
10. Himmelstein, D.S.; Lizee, A.; Hessler, C.; Brueggeman, L.; Chen, S.L.; Hadley, D.; Green, A.; Khankhanian, P.; Baranzini, S.E. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **2017**, *6*, e26726. [CrossRef]
11. Callahan, T.J.; Tripodi, I.J.; Hunter, L.E.; Baumgartner, W.A. *A Framework for Automated Construction of Heterogeneous Large-Scale Biomedical Knowledge Graphs*; Technical Report; Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: Article; Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor, NY, USA, 2020. [CrossRef]
12. Köhler, S.; Gargano, M.; Matentzoglu, N.; Carmody, L.C.; Lewis-Smith, D.; Vasilevsky, N.A.; Danis, D.; Balagura, G.; Baynam, G.; Brower, A.M.; et al. The human phenotype ontology in 2021. *Nucleic Acids Res.* **2020**, *49*, D1207–D1217. [CrossRef] [PubMed]
13. Huntley, R.P.; Harris, M.A.; Alam-Faruque, Y.; Blake, J.A.; Carbon, S.; Dietze, H.; Dimmer, E.C.; Foulger, R.E.; Hill, D.P.; Khodiyar, V.K.; et al. A method for increasing expressivity of Gene Ontology annotations using a compositional approach. *BMC Bioinform.* **2014**, *15*, 155. [CrossRef] [PubMed]
14. Wishart, D.; Arndt, D.; Pon, A.; Sajed, T.; Guo, A.C.; Djoumbou, Y.; Knox, C.; Wilson, M.; Liang, Y.; Grant, J.; et al. T3DB: The toxic exposome database. *Nucleic Acids Res.* **2015**, *43*, D928–934. [CrossRef] [PubMed]
15. Shefchek, K.A.; Harris, N.L.; Gargano, M.; Matentzoglu, N.; Unni, D.; Brush, M.; Keith, D.; Conlin, T.; Vasilevsky, N.; Zhang, X.A.; et al. The Monarch Initiative in 2019: An integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* **2020**, *48*, D704–D715. [CrossRef] [PubMed]
16. Landrum, M.J.; Chitipiralla, S.; Brown, G.R.; Chen, C.; Gu, B.; Hart, J.; Hoffman, D.; Jang, W.; Kaur, K.; Liu, C.; et al. ClinVar: Improvements to accessing data. *Nucleic Acids Res.* **2020**, *48*, D835–D844. [CrossRef]
17. Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **2016**, *44*, D1214–D1219. [CrossRef]
18. Natale, D.A.; Arighi, C.N.; Blake, J.A.; Bona, J.; Chen, C.; Chen, S.C.; Christie, K.R.; Cowart, J.; D'Eustachio, P.; Diehl, A.D.; et al. Protein Ontology (PRO): Enhancing and scaling up the representation of protein entities. *Nucleic Acids Res.* **2017**, *45*, D339–D346. [CrossRef]
19. Diehl, A.D.; Meehan, T.F.; Bradford, Y.M.; Brush, M.H.; Dahdul, W.M.; Dougall, D.S.; He, Y.; Osumi-Sutherland, D.; Ruttenberg, A.; Sarntivijai, S.; et al. The Cell Ontology 2016: Enhanced content, modularization, and ontology interoperability. *J. Biomed. Semant.* **2016**, *7*, 44. [CrossRef]

20. AOP-Wiki. Available online: https://aopwiki.org (accessed on 18 October 2021).
21. Barbie, D.A.; Tamayo, P.; Boehm, J.S.; Kim, S.Y.; Moody, S.E.; Dunn, I.F.; Schinzel, A.C.; Sandy, P.; Meylan, E.; Scholl, C.; et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **2009**, *462*, 108–112. [CrossRef]
22. Affò, S.; Dominguez, M.; Lozano, J.J.; Sancho-Bru, P.; Rodrigo-Torres, D.; Morales-Ibanez, O.; Moreno, M.; Millán, C.; Loaeza-del Castillo, A.; Altamirano, J.; et al. Transcriptome analysis identifies TNF superfamily receptors as potential therapeutic targets in alcoholic hepatitis. *Gut* **2013**, *62*, 452–460. [CrossRef]
23. Neusser, M.A.; Lindenmeyer, M.T.; Moll, A.G.; Segerer, S.; Edenhofer, I.; Sen, K.; Stiehl, D.P.; Kretzler, M.; Gröne, H.J.; Schlöndorff, D.; et al. Human nephrosclerosis triggers a hypoxia-related glomerulopathy. *Am. J. Pathol.* **2010**, *176*, 594–607. [CrossRef]
24. Fang, Z. GSEApy, 2022. original-date: 2016-01-09T03:05:06Z. Available online: https://github.com/zqfang/GSEApy (accessed on 20 July 2022).
25. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef] [PubMed]
26. Saito, M.; Chakraborty, G.; Mao, R.F.; Wang, R.; Cooper, T.B.; Vadasz, C.; Saito, M. Ethanol alters lipid profiles and phosphorylation status of AMP-activated protein kinase in the neonatal mouse brain. *J. Neurochem.* **2007**, *103*, 1208–1218. [CrossRef] [PubMed]