scientific reports

OPEN

# Hypothesizing mechanistic links between microbes and disease using knowledge graphs

Brook E. Santangelo[1]✉, Michael Bada[2], Lawrence E. Hunter[2] & Catherine Lozupone[1]

**Knowledge graphs have been a useful tool for many biomedical applications because of their effective representation of biological concepts. Plentiful evidence exists linking the gut microbiome to disease in a correlative context, but uncovering the mechanistic explanation for those associations remains a challenge. Here we demonstrate the potential of knowledge graphs to hypothesize plausible mechanistic accounts of host-microbe interactions in disease. We have constructed a knowledge graph of linked microbes, genes and metabolites called MGMLink, and, using a shortest path or template-based search through the graph and a novel path-prioritization methodology based on the structure of the knowledge graph, we show that this knowledge supports inference of mechanistic hypotheses that explain observed relationships between microbes and disease phenotypes. We discuss specific applications of this methodology in inflammatory bowel disease and Parkinson's disease. This approach enables mechanistic hypotheses surrounding the complex interactions between gut microbes and disease to be generated in a scalable and comprehensive manner.**

**Keywords** Knowledge graphs, Embedding methods, Microbiome, Computational biology

That gut microbiome composition differs in disease states has become increasingly clear from metagenomic studies[1,2]. Multi-omic analyses are critical to our understanding of how the microbial environment influences host metabolism or genetics in various disease contexts[3,4]. With more than 20,000 citations relating the microbiome to disease now published each year in PubMed since 2019, there is immense potential for integrating existing studies and knowledge into a centralized knowledge base[5]. Knowledge graphs (KGs) describe relationships between entities of different types, *e.g.,* how a gene or gene product influences a disease, or how a metabolite is processed in a specific pathway. KGs serve many purposes, from more effective search of biological relationships in a specific context to predicting new relationships based on the graph structure. KGs have broad applications across biomedical research, including prediction of drug-drug interactions[6,7], evaluation of mechanisms of toxicity[8], prediction of unknown drug disease targets[9], and linking symptoms from electronic health records to better understand disease[10]. Recent advances in the development of microbially relevant KGs have shown promise in aiding our understanding of the role of microorganisms in many environments. One outcome of KGs is to utilize the integrated knowledge to extract causal explanations of an underlying phenomenon, termed mechanistic inference. Existing methods of mechanistic inference using KGs have been applied to drug discovery meta path-based constraint or embeddings-based node similarity, but generally the resulting paths are short, suggesting associations between entities rather than mechanistic explanations[7,8,11].

Our understanding of specific microbiome signatures on clinical outcomes is sparse and mainly correlative[12,13]. That is, microbiome signatures have been associated with many diseases including autoimmune, gastrointestinal, cancer, and neurological disease[14,15], resulting in correlative conclusions that often lack a mechanistic account. Experimental studies to test hypothesized mechanisms are critical, but they are expensive and cannot cover all potential pathways a microbe might act on within a host. The generation of robust mechanistic hypotheses mediating host-microbe interactions that merit the expense of experimental follow-up relies on integration of novel observations with existing knowledge. This process is typically manual and time-consuming, requiring investigators to comb through unstructured text and disparate databases that document previous findings. Without a mechanistic account we do not fully understand the role of the gut microbiome in disease, which makes the development of better and more targeted therapeutics difficult. The applications of mechanistic inference over KGs have yet to be extended into the microbiome field[5,16]. Additionally, existing

[1]Department of Biomedical Informatics, University of Colorado Denver Anschutz Medical Campus, Aurora, CO, USA. [2]Department of Pediatrics, University of Chicago, Chicago, IL, USA. ✉email: brook.santangelo@cuanschutz.edu

microbial KGs incorporate information more focused on microbial trait outcomes[17] or are limited in scope and/or lacking in biomedical information about the host[2,18,19].

To augment the ability to use existing knowledge from prior studies to generate mechanistic hypotheses, we present a methodology that integrates microbiome information into a KG and generates hypothetical mechanistic accounts of how microbes influence disease. We created a microbiome-relevant KG representing microbe-gene-metabolite links called MGMLink (Microbe-Gene-Metabolite-Link) by integrating known microbe-host interactions into a biomedical KG built using the PheKnowLator framework (Fig. 1)[20]. To augment the default PheKnowLator KG with information on microbes, we integrated data from gutMGene, a manually curated repository of assertions involving gut microbes, microbial metabolites, and target genes from over 360 PubMed publications[21]. We also introduce two methods for examining mechanisms that describe the interaction between a microbe and a target entity though paths in the graph, one using the graph structure and one using semantic constraints. We demonstrate the utility of this KG in the discovery of microbial mechanisms by exploring the extent to which microbes included in the KG are relevant in specific diseases (inflammatory bowel disease and Parkinson's disease) and illustrate feasible mechanisms of action between microbes and these diseases. We thus show that the combination of knowledge from gutMGene and PheKnowLator and these path search methodologies facilitate hypothesis generation regarding mechanisms linking microbes with disease.

## Methods

### Knowledge representation and incorporation

To generate MGMLink, we incorporated previously published microbe-host interactions from the gutMGene database into the PheKnowLator framework[20,21]. PheKnowLator is a Python 3 library that enables construction of KGs that incorporate a wide variety of data and terminology sources, including ontologies such as the Mondo Disease Ontology (MONDO), the Chemical Entities of Biological Interest Ontology (CHEBI), and the Human Phenotype Ontology (HPO)[20]. The framework allows alternative knowledge modeling approaches; we used a model in which concepts are unidirectionally relationally linked in the graph (as opposed to bidirectionally)[20]. We also use a version of a PheKnowLator KG that reflects a transformation based on OWL-NETS (Network Transforms for Statistical learning), which enables network inference of Web Ontology Language (OWL)-encoded knowledge via abstraction into biologically meaningful triples[22]. These parameters produce the topologically simplest graph and have the best performance in relevant metrics such as node embedding quality. The representation of gutMGene data in the framework was matched to the PheKnowLator framework parameters. All assertions were mapped to an OWL-encoded KG representation and then transformed into triples using OWL-NETS, resulting in four unique patterns, examples of which are shown in Table 1 [22]. This resulted in new KG nodes that represented microbes in the context of the anatomical location (*i.e.,* the gut) and species (human or mouse) in which the interactions have been reported to occur.

The gutMGene database consists of microbe-metabolite and microbe-gene assertions that occur in the host of either human or mouse[21]. These relationships were manually extracted from over 360 PubMed publications and are based on validated methods such as RT-qPCR, high-performance liquid chromatography, and 16S rRNA sequencing[21]. Specifically, the gutMGene database describes four different types of assertions: (1) microbial substrates observed in humans or mice, (2) microbial metabolites observed in humans or mice, (3) genes observed to be inhibited by microbes in humans and mice, and (4) genes observed to be activated by microbes in humans and mice[21]. Assertions 1 and 2 were extracted from the Association between Gut microbe and Metabolite v1.0 results for Human and Mouse, respectively, and assertions 3 and 4 were extracted from the Association between Gut microbe and Gene v1.0 for Human and Mouse, respectively. These were then represented using a specific semantic pattern to encompass the relationship type and the context (Table 1). We attempted to map each microbial type to an entry in the NCBI Taxonomy, and for those microbial types that could not be mapped to NCBI Taxonomy entries, new nodes representing unclassified bacterial organisms

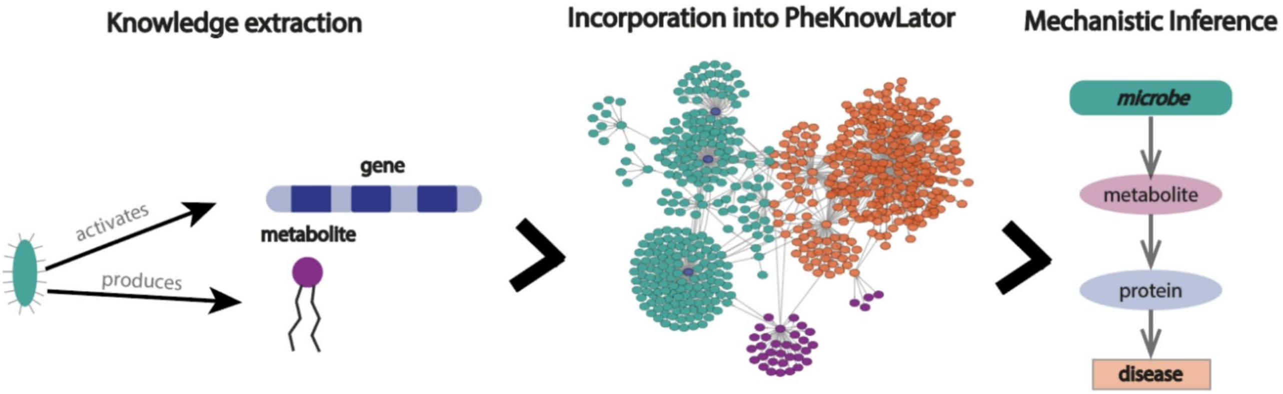


**Fig. 1**. Microbe-host gene and microbe-metabolite relationships such as the types shown (microbe activates gene or microbe produces metabolite) using the gutMGene database were incorporated into the PheKnowLator framework. The microbiome-relevant KG was then used to examine paths between microbes and diseases of interest.

| New KG assertion | OWL – NETS triple representation |
|---|---|
| *Streptococcus* in lower digestive tract in human inhibits CCL20 | St subclass_of NCBITaxon: Streptococcus |
| | St located_in UBERON:'lower digestive tract' |
| | St located_in NCBITaxon:'Homo sapiens' |
| | St indirectly_negatively_regulates_activity_of NCBIGene:CCL20 |
| *A. muciniphila* in lower digestive tract in human activates IL10 | Am subclass_of NCBITaxon:'Akkermansia muciniphila' |
| | Am located_in UBERON:'lower digestive tract' |
| | Am located_in NCBITaxon:'Homo sapiens' |
| | Am indirectly_positively_regulates_activity_of NCBIGene:IL10 |
| *F. prauznitzii* in lower digestive tract in mouse produces benzoic acid | Fp subclass_of NCBITaxon:'Faecalibacterium prauznitzii' |
| | Fp located_in UBERON:'lower digestive tract' |
| | Fp located_in NCBITaxon:'Mus musculus' |
| | Fp contains_process o has_output CHEBI:benzoic acid |
| *R. bromii* in lower digestive tract in human metabolizes propionate | Rb subclass_of NCBITaxon:'Ruminococcus bromii' |
| | Rb located_in UBERON:'lower digestive tract' |
| | Rb located_in NCBITaxon:'Homo sapiens' |
| | Rb involved_in_positive_regulation_of o has_primary_input_or_output CHEBI:propionate |

**Table 1**. Specific examples of all types of patterns identified from gutMGene assertions, and corresponding OWL-NETS representations into which these assertions were converted to integrate them into the knowledge base.

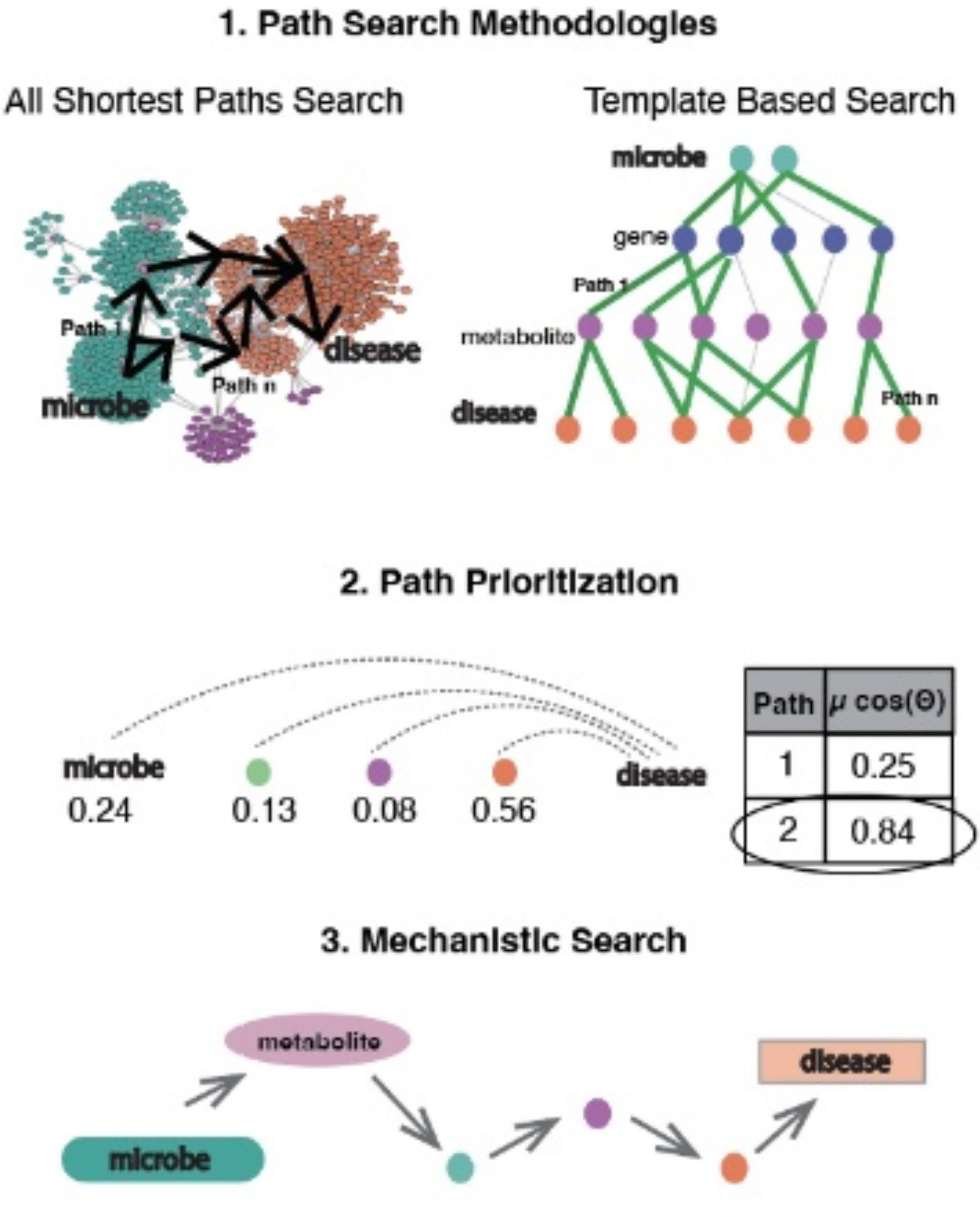
were created. Additionally, microbes that had an unclear or indirect mapping to a NCBITaxon identifier were corrected by introducing edges that capture the taxonomic rank and relationships. A total of 1874 assertions from gutMGene were created in MGMLink for 533 unique microbial taxa (at the family, genus, species, or strain level), which was then added to the 5,072,031 edges and 781,043 nodes. NCBI Taxonomy entries were found for 336 of these taxa, and taxonomic information is incorporated for 461 of the gutMGene assertions. To generate a connected graph, species or strains were related to their corresponding higher-level classifications (genus, family, class, etc.) which already existed in the KG. This allowed for inferences to be made for a microbe unmapped to NBCI Taxonomy based on species-strain or genus-species relationships. In total, the MGMLink KG has 5,076,297 edges and 782,466 nodes. The framework to build MGMLink is available as a SnakeMake workflow in a github repository, with all relevant input data[23].

## Mechanism prediction framework

The path between two nodes in a KG can be drawn in a nearly infinite number of ways, especially for a KG of this size (over 780,000 nodes and 5,000,000 edges). We employ two path search methodologies: an all shortest path and a template-based search. We found all shortest paths using a breadth-first search algorithm with unweighted edges, which incrementally searches all neighbors of a source node until the target node is found and returns the paths with the minimum number of edges. Shortest path search is a common approach to examining a mechanistic link between two nodes, however the randomness can introduce uninteresting and irrelevant nodes[24,25]. This unbiased methodology can still allow for a comprehensive assessment of potential interactions between microbes and metabolites, proteins, or processes of interest. We produced shortest paths in both a directed and undirected manner such that paths may not always flow in one direction. The source input into this process was the first-order neighbor of a microbe. We used the microbial first-order neighbors which represented the anatomical location and species of interest for a given microbe (e.g., 'streptococcus: lower digestive tract Homo sapiens as shown in Table 1). This constraint ensured that paths would include the microbial relationships to host genes or microbial metabolites introduced from gutMGene according to the semantic model described in Table 1.

Next, we introduce a template-based search methodology. This method allows us to traverse the graph using a specific set of semantic constraints on node type, in this case based on ontology prefix. Template-based search in KGs has been applied in multiple drug or disease contexts. Also known as a meta-path search, defining a specific set and order of entity and relation types has been applied to predict drug-target interactions, microRNA (miRNA)-disease associations, or compound-protein interactions[11,26,27]. The template-based search method supports the traversal of longer, more mechanistically relevant paths within a microbial KG beyond node–node associations. The template-based search generates a series of tables representing all connected nodes according to the given template and traces the paths that exist with members of the specified node types at each position. For example, in the template 'microbe, metabolite, protein, gene, disease', we searched for microbe-metabolite pairs, metabolite-protein pairs, protein-gene pairs, and gene-disease pairs and connected all overlapping nodes to form a path. Due to strict constraints that are already imposed from the template-based search, we performed this in an undirected manner. For the template-based search we did not constrain the type of edge between each node in the path, though this is supported within the framework.

There are frequently many ties among shortest paths and template-based paths, so we devised a novel metric to prioritize among these paths. The metric considers the structure of the graph by using graph embeddings. For each path, the cosine similarity is calculated between all nodes (source node and all intermediate nodes) in the path and the target node. We then rank paths based on their average cosine similarity (Fig. 2). Vector embeddings of MGMLink were generated using Node2Vec and TransE, two common graph embeddings methods, for comparison[28,29]. By maximizing average cosine similarity among paths, the path search is biased towards nodes that are most similar to the target node, which we hypothesized would prioritize more relevant paths and allow the user to select interesting and biologically useful paths based on this ranking system (Fig. 2).



| Path | μ cos(Θ) |
|---|---|
| 1 | 0.25 |
| 2 | 0.84 |

**Fig. 2**. Methodology used to prioritize paths between a given microbe and disease of interest.

## Results

To assess the quality of MGMLink and the prioritized path-based approach to hypothesizing mechanisms, we explored two representative diseases with known microbial mechanisms, inflammatory bowel disease (IBD) and Parkinson's disease (PD). Many studies have identified changes in the microbiome for individuals with IBD compared to healthy individuals, citing significantly decreased abundances of Prevotella, Faecalibacterium, Clostridium, Bifidobacterium and increased abundances of Fusobacterium and Gardnerella, among others[30–32]. However, we are in the early stages of exploring the mechanisms of microbial influence in IBD[33–36]. PD is the second most prevalent neurological disease and is associated with motor, gastrointestinal, and psychiatric dysfunction[37]. It is unknown whether PD starts in the gut, though integrated understanding of intestinal permeability and the pathological implications of the disease highlight the importance of the gut microbiome in the development of PD. Bacteria can produce neurotransmitters and neuromodulators that may affect the pathogenesis of the disease, and the differences in microbiome signatures of individuals with PD compared to healthy controls have become more clear[18]. We evaluated all shortest paths and template-based paths between microbes and IBD or PD. As shown in Fig. 3, there are an abundance of paths between microbes and IBD or PD in MGMLink. This finding reiterates the importance of our methodology to prioritize all shortest paths between two nodes of interest.

### Inflammatory bowel disease

We next aimed to determine whether we could find putative mechanistic paths for microbes previously linked with IBD using MGMLink. We applied the shortest path and template-based methodologies to a study that compared individuals with and without IBD using host gene expression and 16S ribosomal RNA (rRNA)-based estimates of microbial relative[38]. This allowed us to determine whether MGMLink could provide plausible mechanistic paths between microbe-IBD relationships than would be evident from this study alone. IBD is characterized by chronic inflammation of the gut, where excess cytokine production in the mucosa results in gastrointestinal discomfort[39]. Generally, the pathogenesis of IBD is thought to be driven by dysbiosis of the microbiome that influences an abnormal immune response, though the direct cause of key microbes in the

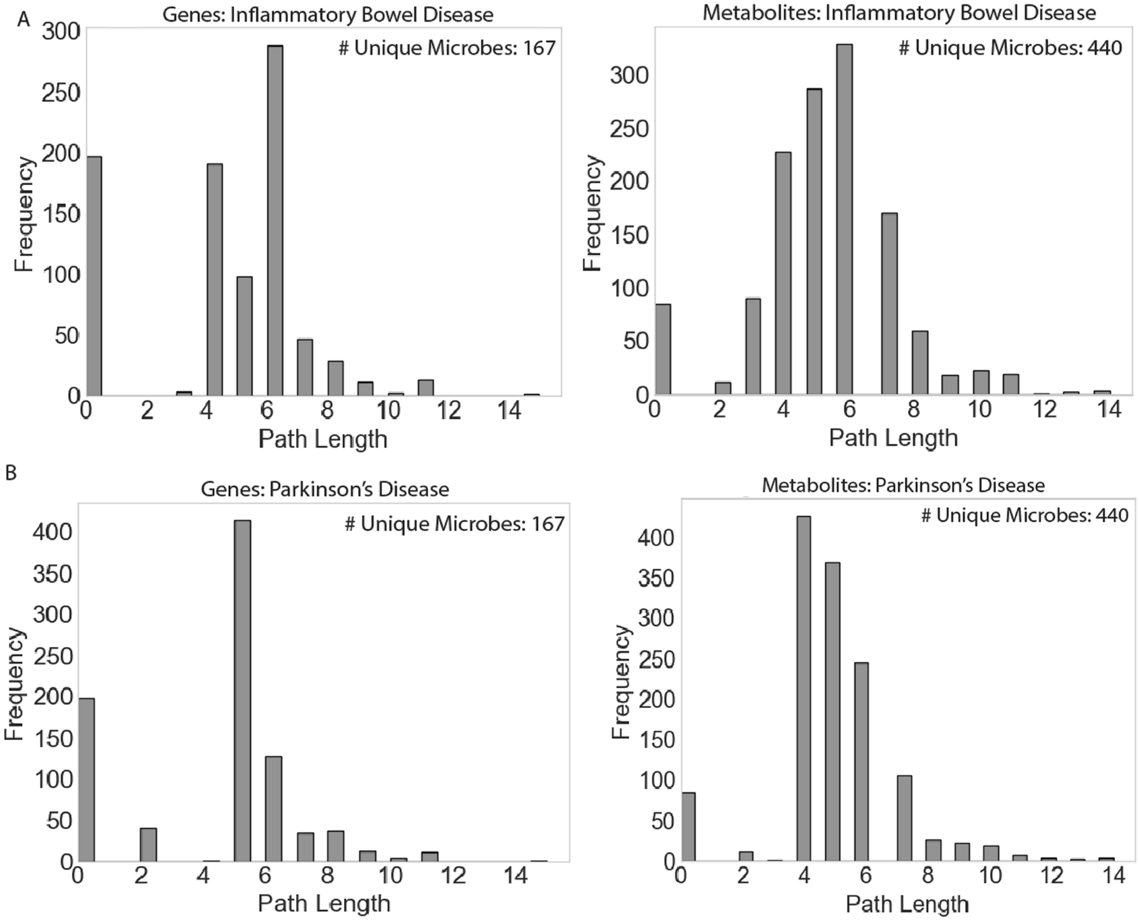


**Fig. 3**. All genes or metabolites that were found to have a potential path through the graph to IBD (**A**) or PD (**B**). The left panel shows the frequency of path lengths that went through a microbe-host gene interaction, and the right through a microbe-metabolite interaction. A path length of 0 represents a case where no path could be found between an entity and the corresponding disease. The number of microbes from which a path could be drawn to the corresponding disease through a gene or metabolite is also reported.

disease is unknown[33]. We examined the paths between microbes that differed significantly between IBD and healthy controls in Lloyd Price et al. to hypothesize potential mechanisms underlying differentially abundant microbes in IBD and the pathogenesis of the disease[38].

The multi-omic study, conducted as part of the Integrative Human Microbiome Project, obtained host and microbial gene expression and 16S-rRNA based microbiome signatures of 132 individuals with or without IBD over the course of one year through sampling of stool, biopsy, and blood[38]. Significantly differentially expressed genes identified from biopsies in the ileum and rectum were compared to differentially abundant microbes identified through 16S sequencing of identical biopsy samples based on partial Spearman correlation that accounted for BMI, age, sex, and diagnosis. In total, 178 gene-microbe pairs that co-varied and were significantly different among IBD and non-IBD individuals were identified[38]. The gutMGene database only contained 12 out of 40 represented microbes from the gene-microbe pairs, alluding to the incomplete nature of MGMLink which limits its application. Three of the 178 gene-microbe pairs existed in MGMLink allowing for further interpretation of the suggested microbe-host interaction (Table 2). These microbe-gene relationships that exist in gutMGene were identified in the same IBD analysis over which this evaluation was done. Although confirmatory evidence of this interaction from other independent sources is lacking in MGMLink, its presence suggests that this knowledge base can make cited studies more accessible. More interestingly, this methodology uncovered potential mechanistic explanations for this gene-microbe-disease interaction that were previously unknown.

We identified all microbial first order neighbors of each of the microbes of interest (*Eubacterium rectale*, *Streptococcus*, and *Eikenella*), and evaluated all shortest and template-based paths between that neighbor and IBD (Supplementary Table 2). We examined the results of *Streptococcus,* which was found to be negatively correlated (R2 = -0.53) with CCL20 in the original published analysis[38], by evaluating paths between the only *Streptococcus* first order neighbor (*Streptococcus:* lower digestive tract Homo sapiens) and IBD. We found a difference in the mechanistic detail between performing a directed verses an undirected all shortest paths-based search. The undirected paths were shorter and often resulted in paths lacking molecular detail necessary to derive a mechanism. The top ranked directed shortest path only elucidated the relationship between the gene CCL20 and ulcerative colitis, a subtype of IBD (Fig. 4). One path of interest was found using both a template-based search (via the template 'microbe, gene, protein, protein, metabolite, disease') and a directed all shortest paths-based search. This path represents a hypothesized mechanism that involved C–C motif chemokine 20, IL-1β, and prostaglandin E synthase (Fig. 4). A previous study cited CCL20 as an inducer of IL-1β, a pro-inflammatory cytokine that is increased in IBD and results in the chronic inflammation that is a staple of the IBD phenotype[33]. This was done by evaluating cytokine signatures of peripheral blood mononuclear cells extracted from individuals with IBD, suggesting a clear association between CCL20 and IL-1β[33]. Another study which used a mouse model of Ulcerative Colitis (UC), a class of IBD, to evaluate the role of prostaglandins, a known inhibitor of the inflammatory response, in the phenotype observed from UC found that prostaglandin E2 (PGE2) plays a significant role in intestinal homeostasis[40]. PGE2 synthesis is mediated by multiple enzymes, including microsomal prostaglandin E synthase, included in this path identified. *Streptococcus pneumoniae* was cited as a pathogenic microbe that can stimulate IL-1β secretion which is further boosted by PGE2 signaling[41]. PGE2 signaling can occur through 4 different G-protein coupled receptors resulting in differing immune signaling degradation into an inactive form. The exact mechanism at play here is unknown[41], however these studies further support the plausible mechanism predicted in Fig. 4. The set of paths generated from this microbe-disease search favors using a directed search for all shortest paths or a template-based search over an undirected shortest path-based search to show specific biological detail necessary for mechanistic hypotheses. Although these results have yet to be validated with an experimental approach, the potential for generating targets in a high throughput manner is clear and this method reduces the search space of microbial targets.

## Parkinson's disease

In the next case study, we identified a microbe of potential interest in PD, *Faecalibacterium prauznitzii*. This microbe had been previously described to have a protective effect for PD[37]. This relationship was then further explored using MGMLink to hypothesize a potential mechanism of the microbe's protective effects. Although there are limitations in this methodology with a biased and incomplete knowledge base, it is promising that we were able to identify interesting paths for a prominent microbe that had already been associated with PD.

The same path search methodologies were conducted between *F. prausnitzii* and PD, again evaluating paths from each microbial first order neighbor of *F. prausnitzii* as the first step in the path (*Faecalibacterium prausnitzii*: lower digestive tract Homo sapiens and *Faecalibacterium prausnitzii*: lower digestive tract Mus musculus). We found that the ranking of paths based on average cosine similarity varied significantly across embeddings methods (Supplementary Table 1). However, when comparing average cosine similarity values for paths found via all shortest paths and template-based search, we found that values were higher for paths found using a template-based search for this microbe-disease pair (Fig. 5A, Supplementary Figure 1). This suggests that a template-based search methodology can find more relevant paths. To further support this claim, we examined

| Microbe | Gene | Regulation |
|---|---|---|
| Eubacterium rectale (human) | CXCL6 | Negative |
| Streptococcus (human) | CCL20 | Negative |
| Eikenella (human) | CCL20 | Negative |

**Table 2.** Gene-microbe pairs from gene expression and microbial abundance analysis that exist in MGMLink.

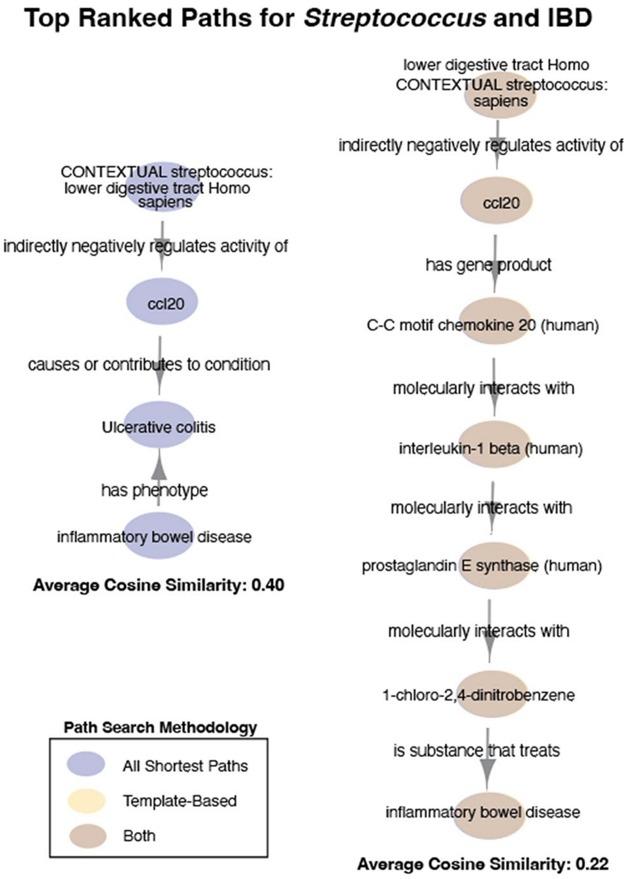

Top Ranked Paths for Streptococcus and IBD
CONTEXTUAL streptococcus: lower digestive tract Homo sapiens
indirectly negatively regulates activity of
ccl20
causes or contributes to condition
Ulcerative colitis
has phenotype
inflammatory bowel disease
Average Cosine Similarity: 0.40
lower digestive tract Homo CONTEXTUAL streptococcus: sapiens
indirectly negatively regulates activity of
ccl20
has gene product
C-C motif chemokine 20 (human)
molecularly interacts with
interleukin-1 beta (human)
molecularly interacts with
prostaglandin E synthase (human)
molecularly interacts with
1-chloro-2,4-dinitrobenzene
is substance that treats
inflammatory bowel disease
Average Cosine Similarity: 0.22
Path Search Methodology
All Shortest Paths
Template-Based
Both

**Fig. 4**. Path search results from searching the microbial first order neighbor of *Streptococcus* as the source and inflammatory bowel disease as the target. Top prioritized path from an undirected shortest path search is shown in blue (left). The path in the mixed yellow and blue color was found using a template-based path search, using the template 'microbe, gene, protein, protein, metabolite, disease', and was also found using a directed shortest path-based search.

low- and high-ranking paths from both search methodologies. The all shortest paths-based search found a path connecting *F. prausnitzii* to PD through the simple fact that both are connected to *Homo sapiens* (Fig. 5B). While true, this uninteresting path illustrates a common trend for shortest path searches that find paths through broad nodes. The highest-ranking path, also a shortest path, does not provide an interesting molecular mechanistic account, as dementia is only a phenotype of PD (Fig. 5B). The template-based search found a mechanistic path suggesting that a metabolite produced by *F. prausnitzii* interacts with the pro-inflammatory cytokine IL-6 (Fig. 5B). Benzoic acid has been seen to have anti-inflammatory effects[42–44], and *F. prausnitzii* has been observed to be decreased in patients with PD. It is plausible that a reduction in the anti-inflammatory microbe *F. prauznitzii* may be implicated in the inflammatory state of a person with PD through the lack of inhibition of the pro-inflammatory cytokine IL-6. This path had a relatively low-ranking average cosine similarity score (0.11) and still has scientific merit. High ranking template-based paths also bring interesting results, as IL-12b is a pro-inflammatory cytokine found to have increased activation in PD, and reactive nitrogen species including nitric oxide have been of interest to treat PD due to their role in neurodegeneration[45]. This comparison of path search methods suggests that semantic constraint to find interesting mechanistic paths performs better than structural methods. Shortest path search finds false positives, such as paths that are true but not mechanistic, across the range of average cosine similarity scores (Fig. 5). The path prioritization method does not sufficiently differentiate the uninteresting paths found using a shortest path-based result. However, mechanistic hypotheses can be found from both high- and low-ranking template-based paths.

## Discussion and conclusion

KGs have extensive applications in the biomedical field because they integrate complex concepts in a systematic way. The knowledge represented in MGMLink expands upon previously observed microbe-gene interactions with mechanistic explanations that can infer biological relationships of interest. We apply these concepts to generate mechanistic hypotheses about microbial influence on diseases such as IBD and PD. The two search methodologies introduced here generate long paths that can serve as mechanistic explanations for the role of the microbiome in disease. The resulting paths found for the *Streptococcus*-IBD pair illustrate how a directed shortest path-based search or a template-based search can find more mechanistically relevant nodes and paths

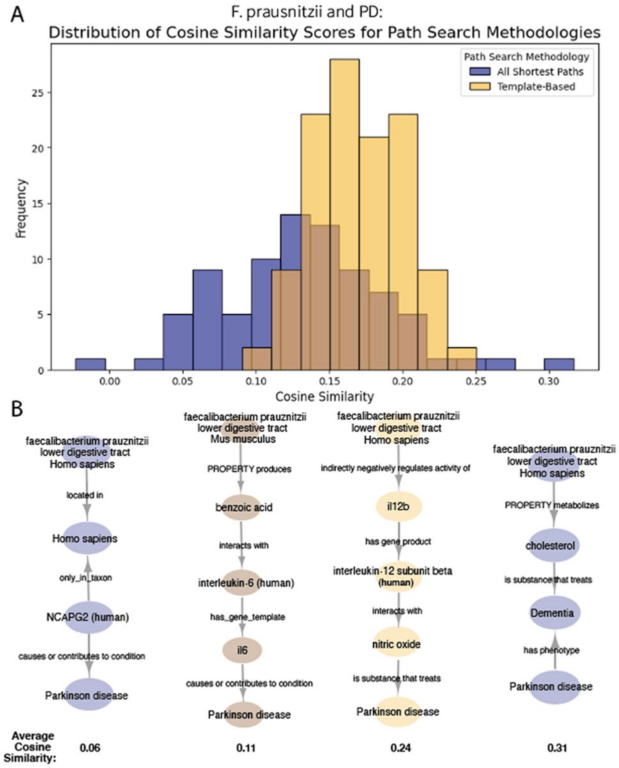


**Fig. 5**. Path prioritization results for searching all microbial first order neighbors of *Faecalibacterium prausnitzii* as the source and PD as the target. (**A**) Cosine similarity values represent the average cosine similarity of all nodes to the target node for each path from embeddings generated using node2vec. A Mann–Whitney U test revealed a significant difference in cosine similarity values between methods ($p = 8.02\text{e-}13$). (**B**) Exemplary paths are shown for all shortest paths (blue) and template-based paths (yellow), where one path was found from both algorithms and is show with both colors. The templates used were 'microbe, metabolite, protein, gene, disease' which applies to the path with the value 0.11 and 'microbe, gene, protein, metabolite, disease' which applies to the path with value 0.31.

than an undirected shortest path-based search. The *F. prausnitzii*-PD pair identified more interesting paths using the template-based method, demonstrating the importance of semantic constraint in mechanistic search. In comparing shortest paths and template-based paths, we found that directed shortest paths and template-based paths introduce more mechanistically interesting paths than undirected shortest paths. These results demonstrate that with an effective semantically constrained search methodology, paths found in MGMLink present feasible descriptions of the ways a microbe may influence disease.

Existing methods of mechanistic inference using KGs include question/answering using graph query languages, link prediction leveraging graph structure, or embeddings based models[7,46]. Previous semantic constraint methods have found shorter paths, such as gene–gene, gene-disease, or drug-disease pairs that guide biological understanding of potential interactions[7,11]. Evidence for mechanistic hypotheses based on these predicted node–node relationships are rarely the forefront of these search methodologies. The all shortest paths-based and template-based methodologies that are introduced here uncover the applications of a more comprehensive mechanistic account using KGs in a microbiome-relevant context. This work introduces methods that provide detailed mechanistic explanations through the identification and prioritization of longer paths in a KG.

In recent years, microbe-relevant KGs have come to light that include knowledge surrounding microbial environments, traits, and in some cases involvement in disease[16]. However existing microbial KGs lack the breadth and depth of knowledge required to support our goals of generating detailed mechanistic hypotheses about disease. For example, the Human Microbe Disease Network (HMDN) was constructed from manual curation of microbe-disease associations across published studies, however without broadly relevant host knowledge, complete mechanistic paths cannot be drawn from this resource[19]. The Microbiome KG was constructed by manually extracting entities and relationships from the supplementary tables of 22 microbiome relevant publications, successfully standardizing experimental results that are otherwise difficult to access. However, the small number of publications included limits the extent to which microbe-host mechanisms can be uncovered[47]. MetagenomicKG integrates metagenomic information from various disease- and function-relevant databases and was used to predict pathogens using a graph neural network model[48]. We argue that the incorporation of more host relevant knowledge can enhance the effectiveness of KGs to be used for mechanistic hypothesis generation.

This work introduces a method for using novel representations of existing knowledge about microbial function to hypothesize mechanisms. However, there are limitations to this work. KG construction methodologies include retrieving information from free text in the literature, from structured databases, or from experimental results. We have demonstrated that the incorporation of more host relevant knowledge can enhance the effectiveness of KGs to be used for mechanistic hypothesis generation. Information retrieval approaches from text can be automated, including named entity recognition (NER) and relation extraction (RE) to identify assertions from manuscripts and map them to ontologies, or involve manual curation as is done with gutMGene[49]. MGMLink utilizes a structured database, gutMGene, to incorporate microbial information into the KG PheKnowLator. This approach limits the resource's ability to stay current with the field, as databases only reflect knowledge available at the time of publication. MGMLink contains only the microbes described in the gutMGene database and does not cover all mechanisms by which microbes influence host genes or consume or produce metabolites. Lastly, the use of literature review as an evaluation of these results does not scale in ways necessary to predict unknown, novel hypotheses. In the future, we foresee these methods proposing hypothesized microbe-host interactions that bring researchers closer to targeted lab experimentation. Despite these limitations, this work offers a novel approach to addressing the pressing need to understand the mechanisms underlying the now well-established relationships between gut microbes and human disease.

Here we describe the creation and application of a KG, MGMLink, of microbe-host interactions toward the automated construction of mechanistic hypotheses regarding microbe-disease relationships. We have shown that novel inferences can come of past results with the search methods applied to MGMLink. There is potential to expand the scientific reach of graph search methodologies into the field of the microbiome, and the microbiome-relevant MGMLink KG described here begins to realize this potential.

## Data availability

The code for the framework to construct MGMLink and all relevant data for the analyses provided in this manuscript is freely available and can be accessed online at https://github.com/bsantan/MGMLink.git. All results can also be found at https://doi.org/10.5281/zenodo.14523102.

Received: 8 October 2024; Accepted: 19 February 2025
Published online: 26 February 2025

## References

1. Falony, G. et al. The human microbiome in health and disease: Hype or hope. *Acta Clin. Belg. Int. J. Clin. Lab. Med.* **74**, 53–64 (2019).
2. King, C. H. et al. Baseline human gut microbiota profile in healthy people and standard reporting template. *PLoS ONE* **14**, e0206484 (2019).
3. Ning, L. et al. Microbiome and metabolome features in inflammatory bowel disease via multi-omics integration analyses across cohorts. *Nat. Commun.* **14**, 7135 (2023).
4. Wallen, Z. D. et al. Metagenomics of Parkinson's disease implicates the gut microbiome in multiple disease mechanisms. *Nat. Commun.* **13**, 6958 (2022).
5. Badal, V. D. et al. Challenges in the construction of knowledge bases for human microbiome-disease associations. *Microbiome* **7**, 129. https://doi.org/10.1186/s40168-019-0742-2 (2019).
6. Nicholson, D. N. & Greene, C. S. Constructing knowledge graphs and their biomedical applications. *Comput. Struct. Biotechnol. J.* **18**, 1414–1428. https://doi.org/10.1016/j.csbj.2020.05.017 (2020).
7. Reese, J. T. et al. KG-COVID-19: A framework to produce customized knowledge graphs for COVID-19 response. *Patterns* **2**, 100155 (2021).
8. Tripodi, I. J. et al. Applying knowledge-driven mechanistic inference to toxicogenomics. *Toxicol. In Vitro* **66**, 104877 (2020).
9. Mayers, M. et al. Design and application of a knowledge network for automatic prioritization of drug mechanisms. *Bioinformatics* **38**, 2880–2891 (2022).
10. Zhang, X. A. et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. *Npj Digit. Med.* **2**, 32 (2019).
11. Himmelstein, D. S. et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **6**, e26726 (2017).
12. Huang, Y. A. et al. Prediction of microbe-disease association from the integration of neighbor and graph with collaborative recommendation model. *J. Transl. Med.* **15**, 209 (2017).
13. Chang, C. S. & Kao, C. Y. Current understanding of the gut microbiota shaping mechanisms. *J. Biomed. Sci.* **26**, 59. https://doi.org/10.1186/s12929-019-0554-5 (2019).
14. Berg, G. et al. Microbiome definition re-visited: old concepts and new challenges. *Microbiome* **8**, 103. https://doi.org/10.1186/s40168-020-00875-0 (2020).
15. Stopińska, K., Radziwoń-Zaleska, M. & Domitrz, I. The microbiota-gut-brain axis as a key to neuropsychiatric disorders: A mini review. *J. Clin. Med.* **10**, 4640. https://doi.org/10.3390/jcm10204640 (2021).
16. Santangelo, B. E. et al. Integrating biological knowledge for mechanistic inference in the host-associated microbiome. *Front. Microbiol.* **15**, 1351678 (2024).
17. Joachimiak, M. P. et al. KG-Microbe: A reference knowledge-graph and platform for harmonized microbial information. In *CEUR Workshop Proceedings* vol. 3073 (2021).
18. Liu, T. et al. Exploring the microbiota-gut-brain axis for mental disorders with knowledge graphs. *J. Artif. Intell. Med. Sci.* **1**, 30–42 (2020).
19. Ma, W. et al. An analysis of human microbe-disease associations. *Brief. Bioinform.* **18**, 85–97 (2017).
20. Callahan, T. J., Tripodi, I. J., Hunter, L. E. & Baumgartner, W. A. A framework for automated construction of heterogeneous large-scale biomedical knowledge graphs. *bioRxiv.* https://doi.org/10.1101/2020.04.30.071407 (2020).
21. Cheng, L. et al. gutMGene: A comprehensive database for target genes of gut microbes and microbial metabolites. *Nucleic Acids Res.* **50**, D795–D800 (2022).
22. Callahan, T. J. et al. OWL-NETS: Transforming OWL representations for improved network inference. In *Pacific Symposium on Biocomputing* (2018).
23. Santangelo, B. E. *MGMLink*. GitHub. https://github.com/bsantan/MGMLink (2024).

24. Zhao, K. et al. Leveraging demonstrations for reinforcement recommendation reasoning over knowledge graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 239–248 (ACM, Virtual Event China, 2020). https://doi.org/10.1145/3397271.3401171.
25. Dijkstra, E. W. A note on two problems in connexion with graphs. *Numer. Math.* **1**, 269–271 (1959).
26. Fu, G. et al. Predicting drug target interactions using meta-path-based semantic network analysis. *BMC Bioinform.* **17**, 160 (2016).
27. Zhang, L. et al. Predicting MiRNA-disease associations by multiple meta-paths fusion graph embedding model. *BMC Bioinform.* **21**, 470 (2020).
28. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 855–864 (ACM, 2016). https://doi.org/10.1145/2939672.2939754.
29. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. & Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems* Vol. 26 (eds Burges, C. J. et al.) (Curran Associates Inc, 2013).
30. Halfvarson, J. et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.* **2**, 17004 (2017).
31. Zhou, Y. et al. Alterations in gut microbial communities across anatomical locations in inflammatory bowel diseases. *Front. Nutr.* **8**, 615064 (2021).
32. Prosberg, M., Bendtsen, F., Vind, I., Petersen, A. M. & Gluud, L. L. The association between the gut microbiota and the inflammatory bowel disease activity: A systematic review and meta-analysis. *Scand. J. Gastroenterol.* **51**, 1407–1415. https://doi.org/10.1080/00365521.2016.1216587 (2016).
33. Skovdahl, H. K. et al. C-C motif ligand 20 (CCL20) and C-C motif chemokine receptor 6 (CCR6) in human peripheral blood mononuclear cells: Dysregulated in ulcerative colitis and a potential role for CCL20 in IL-1β release. *Int. J. Mol. Sci.* **19**, 3257 (2018).
34. Hall, A. B. et al. A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med.* **9**, 103 (2017).
35. Henke, M. T. et al. Capsular polysaccharide correlates with immune response to the human gut microbe Ruminococcus gnavus. *Proc. Natl. Acad. Sci. USA* **118**, e2007595118 (2021).
36. Henke, M. T. et al. Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. *Proc. Natl. Acad. Sci. USA* **116**, 12672–12677 (2019).
37. Hill-Burns, E. M. et al. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. *Mov. Disord.* **32**, 739–749 (2017).
38. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
39. Strober, W., Fuss, I. & Mannon, P. The fundamental basis of inflammatory bowel disease. *J. Clin. Invest.* **117**, 514–521 (2007).
40. Montrose, D. C. et al. The role of PGE2 in intestinal inflammation and tumorigenesis. *Prostaglandins Other Lipid Mediat.* **116–117**, 26–36 (2015).
41. Martínez-Colón, G. J. & Moore, B. B. Prostaglandin E2 as a regulator of immunity to pathogens. *Pharmacol. Therapeut.* **185**, 135–146. https://doi.org/10.1016/j.pharmthera.2017.12.008 (2018).
42. Lee, H. K. et al. Anti-inflammatory effects of OBA-09, a salicylic acid/pyruvate ester, in the postischemic brain. *Brain Res.* **1528**, 68–79 (2013).
43. Zheng, L. T. et al. Inhibition of neuroinflammation by MIF inhibitor 3-({[4-(4-methoxyphenyl)-6-methyl-2-pyrimidinyl] thio}1methyl)benzoic acid (Z-312). *Int. Immunopharmacol.* **98**, 107868 (2021).
44. Tjahjono, Y. et al. Anti-inflammatory activity of 2-((3-(chloromethyl)benzoyl)oxy)benzoic acid in LPS-induced rat model. *Prostaglandins Other Lipid Mediat.* **154**, 106549 (2021).
45. Stykel, M. G. & Ryan, S. D. Nitrosative stress in Parkinson's disease. *Npj Park. Dis.* **8**, 104 (2022).
46. Gao, Z., Ding, P. & Xu, R. KG-Predict: A knowledge graph computational framework for drug repurposing. *J. Biomed. Inform.* **132**, 104133 (2022).
47. Goetz, S. L., Glen, A. K. & Glusman, G. MicrobiomeKG: Bridging microbiome research and host health through knowledge graphs. https://doi.org/10.1101/2024.10.10.617697 (2024).
48. Ma, C., Liu, S. & Koslicki, D. MetagenomicKG: A knowledge graph for metagenomic applications. https://doi.org/10.1101/2024.03.14.585056 (2024).
49. Zhang, Y. et al. BioKG: A comprehensive, large-scale biomedical knowledge graph for AI-powered, data-driven biomedical research. https://doi.org/10.1101/2023.10.13.562216 (2023).

## Acknowledgements

Not applicable

## Author contributions

Conceptualization, B.E.S., L.H., and C.L.; methodology, B.E.S. and M.B.; software, B.E.S.; validation, B.E. S., writing—original draft preparation, B.S.; writing—review and editing, B.S., M.B., L.H., and C.L.; visualization, B.S.; supervision, L.H. and C.L.; All authors have read and agreed to the published version of the manuscript.

## Funding

Research reported in this manuscript was supported by the National Library of Medicine (NLM) of the National Institutes of Health (NIH) under award number T15LM009451. This research was also funded by grant number R01LM013400 and OT2TR003422.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-91230-6.

**Correspondence** and requests for materials should be addressed to B.E.S.