GutBrainIE@CLEF26
Annotation Guidelines

Vanessa Bonato, Giorgio Maria Di Nunzio, Nicola Ferro, Ornella Irrera, Stefano Marchesin, Marco Martinelli, Laura Menotti, Gianmaria Silvello, Federica Vezzani.

Document Structure

Document Structure **1**
1. Overview **3**
2. Annotation Framework **5**
Annotation process in practice 6
3. Annotation Workflow **8**
Documents Retrieval from PubMed 8
Pre-annotation for NER and NEL 8
Annotation Guidelines Definition 9
Hands-on Meeting 9
Manual Annotation - First Phase 10
Prof. Vezzani's Students Annotations 10
Pre-annotation Improvement 11
Midway Meeting 11
Manual Annotation - Second Phase 11
Final Meeting 12
Distantly-Supervised Collection Creation 12
Model Guided Curation 12
Deliverables Release 13
4. Entity Annotation Guidelines **14**
4.1. Entity Labels 14
4.2. General Entity Annotation Rules 22
4.3. Entity Linking Rules 29
5. Relation Annotation Guidelines **31**
5.1. Relation Labels 31
5.2. General Relation Annotation Rules 32
6. Annotation Tools **36**
6.1. Automatic Annotation Tools for Entities 36
6.2. Ontology and Database Search Tools 40
6.3. Category-Specific Search Tools 42
7. Bibliography **47**
8. Appendix **49**
Conceptual System 49

Note: Highlighted lines refer to sections containing information useful for the actual manual annotation phases. These are crucial to know and should be refreshed periodically during annotation.

1. Overview

The GutBrainIE@CLEF26 dataset aims to foster the development of Information Extraction (IE) systems that support experts by automatically extracting and linking knowledge from scientific literature, thereby enhancing the understanding of the gut-brain interplay and its role in neurological diseases. Recent evidence suggests a connection between neurological and gut disorders that might play a critical role in mental health-related conditions such as Multiple Sclerosis, Parkinson's, and Alzheimer's [7-9].

These guidelines aim to assist annotators in consistently labeling named entities and relationships within this dataset. **Named entity labeling** involves identifying and classifying specific text spans (entity mentions) into predefined categories and linking them to the corresponding concept in reference biomedical resources. **Relationship labeling** determines if a particular relationship defined between two entity types holds or not [1]. In cases where multiple relationships are defined between two entity types, relationship labeling also determines which specific relation holds.

The dataset will focus on biomedical titles and abstracts related to the gut microbiota and its effects on major neurological diseases, including Alzheimer's, Parkinson's, multiple sclerosis, and amyotrophic lateral sclerosis [7], extracted from PubMed. Entities and relations of interest are defined by the conceptual system in the Appendix Section.

Formally, the GutBrainIE task is divided into two subtasks:

1. **Named Entity Recognition (NER)**: Participants are provided with PubMed abstracts discussing the gut-brain interplay and are asked to extract named entities related to the defined categories of interest. These entities should then be linked to their corresponding concepts in biomedical reference resources.

2. **Relation Extraction (RE)**: Participants are tasked with identifying relations between pairs of extracted entities within a document (title + abstract).

The submitted results are evaluated based on Precision, Recall, and F1 measures for each subtask, using gold annotations created by a team of expert annotators.

These guidelines build upon the methodologies and practices used in crafting EnzChemRED [1], BioRED [2], and BioASQ-QA [3], providing a common approach for annotating abstracts with minimal ambiguity and the highest possible inter-annotator agreement (IAA).

2. Annotation Framework

The annotation process for the *GutBrainIE* dataset involves three tasks [2]:

1. **Named Entity Recognition (NER)**: Identify and classify text spans into one of the defined entity categories.

   Formally speaking, the goal of NER is to identify mentions of the defined entity categories in the text. The NER task can be seen as one of sequence labeling: text is represented as a sequence of tokens (in our case, words) $x = (x_1, x_2, \cdots, x_n)$ where n denotes the length of the text, and the goal is to classify a sequence of tokens $(x_i, x_{i+1}, \cdots, x_{i+j})$ with $i, (i + j) \in [0, n]$ into a corresponding category $y_i \in (y_1, y_2, \cdots, y_m)$. The set $(y_1, y_2, \cdots, y_m)$ is defined as the label set Y for the model, and each label represents a specific entity category that can be found in texts.

2. **Named Entity Linking (NEL)**: Link each identified entity mention to its corresponding concept in a biomedical reference resource, following the priority order defined for each entity category. This step resolves lexical ambiguities and ensures that each mention is associated with the most specific and semantically accurate concept available.

3. **Relation Extraction (RE)**: Identify relationships between entities, as explicitly or implicitly expressed in the text. This problem comprises two different but complementary tasks:

   - **Binary RE**: Given a pair of entity mentions (e_1, e_2) having labels (y_1, y_2) assuming there is a relation $r \in R$ defined from y_1 to y_2 or vice versa, where R is the set of relations defined for the *GutBrainIE* dataset, the objective is to state whether that relation holds between these two entities. The classification should be consistent with the predefined relation types R defined for the *GutBrainIE* dataset [1].

- **Ternary RE**: Given a pair of entity mentions (e_1, e_2) having labels (y_1, y_2), the objective is to determine if there is a relation $r \in R$ (predicate) that links the ternary tuple (e_1, r, e_2) or (e_2, r, e_1). The entity appearing before r in the ternary tuple is referred to as "head" or "subject", while the one appearing after r is referred to as "tail" or "object". The relation must align with the context in which the entities appear and should fit the predefined relation types R defined for the *GutBrainIE* dataset [1].

Annotation process in practice

The documents are uploaded on the annotation platform *MetaTron* [4] with pre-annotations for NER and NEL to speed up the annotation process [1]. These annotations have been performed by unsupervised algorithms.

Before starting to annotate, annotators are required to **read the guidelines carefully**, paying particular attention to the defined entity and relation labels to ensure a comprehensive understanding and consistency.

We are aware that the defined annotation guidelines may impose certain restrictions that could result in the loss of potentially relevant or important information. However, these rules are crucial for minimizing uncertainties among annotators and, consequently, ensuring more uniform and consistent annotations.

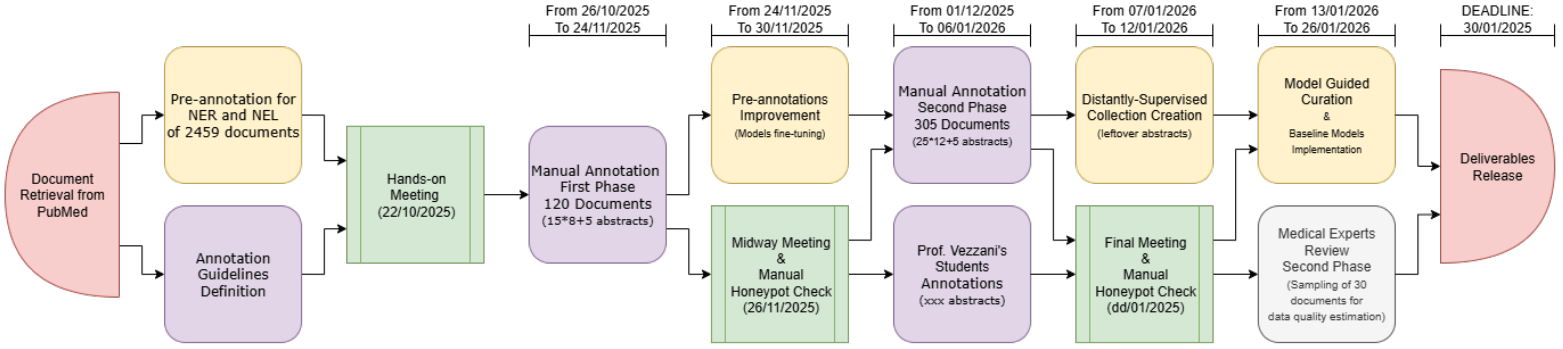
Annotators are required to follow a specific process when annotating each document to ensure consistency and accuracy:

1. **Initial Reading**: First, read carefully through the entire abstract without the pre-annotated mentions visualized. This helps in understanding the context without bias from pre-annotated labels [5].

2. **NER and NEL Annotation:** Once familiar with the abstract, enable the visualization of the pre-annotated entity mentions and proceed with the annotation of named entities and their corresponding links. During this phase, annotators should verify, correct, or complete the entity spans, the assigned categories (NER), and their links to the appropriate biomedical concepts (NEL), following the defined priority order for each entity category. In cases where multiple valid concepts are available across different resources, the concept from the highest-priority resource corresponding to the entity's category should be selected. When an entity instance that appears multiple times in the document is annotated for the first time, you may use the "*annotate all*" feature in MetaTron to uniformly annotate all subsequent occurrences of the same entity mention, ensuring consistent labeling and linking throughout the document.
3. **RE Annotation**: After completing NER, proceed with RE to identify and classify relationships between the annotated entity mentions.

3. Annotation Workflow

The annotation workflow for the *GutBrainIE* dataset is designed to ensure that all annotators have a clear understanding of the steps involved in the annotation process, from document retrieval to the final review of the manually annotated collection. The workflow includes several stages, such as pre-annotation, hands-on meetings, manual annotation phases, model fine-tuning, and medical experts' reviews. The image below details the entire workflow, followed by a step-by-step description of each involved stage.



Documents Retrieval from PubMed

To build the *GutBrainIE* dataset, we selected papers from the PubMed Electronic Database running the following query:

1. ((gut) AND (microbiota OR microbiome)) AND (dementia OR Alzheimer OR multiple sclerosis OR Amyotrophic lateral sclerosis)

Retrieved documents already included in the 2025 collection were removed, resulting in a total of 2458 unique documents.

Pre-annotation for NER and NEL

The retrieved and filtered documents are pre-annotated for NER using the top-performing system among those submitted by participants in the CLEF 2025 challenge. NEL pre-annotations are generated using our baseline linking system. The decision was made

not to pre-annotate the documents for Relation Extraction since even the best CLEF 2025 systems proved insufficiently robust and reliable for this task, making the likelihood of introducing noise significantly higher than that of adding valid relations. Excessive noise in pre-annotations could lead to biases among annotators, ultimately impacting the quality of the final annotated dataset [6].

Annotation Guidelines Definition

The annotation guidelines form the foundation of the *GutBrainIE* dataset's consistency and quality, providing clear instructions on labeling entities and relationships, handling edge cases, and offering examples of typical annotations. Annotators are required to thoroughly familiarize themselves with the guidelines before starting and are encouraged to consult them regularly to ensure consistency, minimize ambiguity, and address uncertainties.

Once the annotation guidelines are finalized, they remain unchanged during the first annotation phase to maintain consistency across annotations. However, a re-validation might be conducted between the first and second annotation phases. In that case, annotations completed during the first phase will be adjusted to reflect these changes, ensuring uniformity across the entire dataset.

Hands-on Meeting

The annotators team conducts a hands-on meeting to exhaustively review the annotation guidelines, familiarize themselves with the annotation tasks, and discuss any potential doubts or ambiguities. During this meeting, annotators jointly annotate several abstracts for NER, NEL, and RE. This collaborative exercise ensures that all annotators are on the same page and understand the expectations clearly [2].

Any uncertainties must be resolved during this session, as annotators won't be allowed to consult with each other during the actual annotation phase. This restriction is intended to prevent the introduction of biases in annotators, which could otherwise compromise the quality and consistency of the annotations.

Manual Annotation - First Phase

In the first phase of manual annotation, **each of the eight annotators will be assigned 15 distinct documents, along with 5 shared documents that form the first half of the honeypot set**. These shared documents are used to evaluate consistency among annotators and to identify any discrepancies in annotation practices. At the end of this phase, a total of 120 uniquely annotated documents plus 5 honeypot documents will be completed.

The involvement of multiple annotators is intended to prevent the models trained on this data from being biased towards the annotation style or task interpretation of a single individual, ensuring a balanced representation of annotation perspectives [7].

During this phase, annotators might note down the most critical doubts or challenges they face, which will later be discussed during the midway meeting.

Prof. Vezzani's Students Annotations

Prof. Vezzani's students from the Translation-oriented terminography course will annotate a separate set of XXX documents, following the same guidelines as the main annotators team. Students will receive YY unique documents each, plus ZZ document belonging to the honeypot set of the main annotators team. The latter will be used to evaluate the performance of the annotators and, therefore, their reliability.

The annotations provided by Prof. Vezzani's students will form a collection named "*Supplementary Annotated Dataset* ". Unlike the "*Gold Standard*" collection, which will be used as the primary training and evaluation resource for the participants of the *GutBrainIE* challenge, this supplementary dataset will serve as an additional resource for further analysis and potential model fine-tuning. This separation ensures the integrity and quality of the *Gold Standard* while providing valuable annotated data from a linguistic perspective.

Pre-annotation Improvement

The annotated data from the first manual annotation phase is used to fine-tune the pre-annotation models to improve the quality and accuracy of the subsequent pre-annotations, reducing the amount of correction needed by annotators in later phases and ensuring a more efficient workflow [1,2].

Midway Meeting

After the first manual annotation phase, the annotators team holds a midway meeting to review progress, discuss critical challenges noted during the annotation phase, and make any necessary adjustments to the guidelines.

During this meeting, the team also evaluates the curated data from the honeypot documents and selects the first sample to be reviewed by medical experts [1]. The goal is to achieve an inter-annotator agreement (IAA) of at least xx% for NER and xx% for RE on the honeypot set [2], ensuring high-quality and consistent annotations across the dataset.

Manual Annotation - Second Phase

In the second manual annotation phase, **each of the eight annotators is assigned 25 distinct documents along with 5 additional shared documents, which form the second half of the honeypot set**. By the end of this phase, the team will have annotated a total of

200 distinct documents and 5 honeypot ones. Combined with the 120 documents from the first phase, the final dataset will contain 320 manually annotated documents.

Final Meeting

Following the second phase of manual annotation, the annotators team holds a final meeting to finalize the annotated collection and address any unresolved issues. During this meeting, annotators will also evaluate the curated data from the honeypot documents and select a second sample of annotated documents for review by medical experts.

Distantly-Supervised Collection Creation

Once the entire set of manually annotated abstracts is completed and approved, the pre-annotation models are fine-tuned again. Specifically, the NER and NEL models are fine-tuned with the new high-quality annotations to further improve its performance. At the same time, we introduce and fine-tune the RE model, *ATLOP*, using the manually annotated relationships, allowing us to use it reliably for Relation Extraction in the next phase.

The newly fine-tuned models are then used to create a distantly-supervised collection by annotating the remaining documents from the original retrieval that were not manually annotated by the main annotation team or by Prof. Vezzani's students. This resulting "*Distantly-Supervised Annotated Dataset*" complements the *Gold Standard* and *Supplementary Annotated* datasets, providing additional training resources while distinguishing between human-validated and automatically generated annotations.

Model Guided Curation

The fine-tuned models generate new predictions for NER and RE, facilitating the identification of potential false positives and negatives in the manually annotated data. By comparing model predictions with human-annotated data, annotators can refine and

re-curate the annotations to further improve the quality of the dataset. Furthermore, this iterative curation process helps in enhancing both the annotated dataset and the model itself, enabling adjustments to model hyperparameters that will be used as baselines for the challenge [1].

Deliverables Release

The deliverables released at the end of the annotation and curation process include:

- **Data Collection**: The annotated dataset will be divided into training, validation, and test sets to facilitate model development and evaluation.
- **Annotation Guidelines**: Detailed documentation including definitions and examples of the entities and relations annotated, as well as the rules followed during the manual annotation phases, to ensure transparency and reproducibility.
- **Web Page for BioASQ**: A dedicated webpage will be set up to allow users to download the corpus and guidelines, with information about how we expect participants to submit their results, along with deadlines for submission.
- **Baseline Models and Performance Metrics**: baseline models and their performances will be made available to participants. We might open Codabench competitions after the deadline for run submissions, allowing participants to further test their predictions on the test set.
- To infinity… and beyond 

4. Entity Annotation Guidelines

4.1. Entity Categories

See the Appendix for a graphical representation of the schema (entities and relationships).

Entity Category	URI	Definition	Reference URL	Notes	Examples
Anatomical Location	NCIT_C13717	Named locations of or within the body. --- The categories "anatomic site" of NCIT and "organism subdivision" of UBERON can be used as references for this label.	link	Instances of "... *axis*" (e.g., "*gut-brain axis*") and "... *system*" (e.g., "*immune system*") should **NOT** be labeled as *anatomical location* entity mentions.	The microbiota resides in various parts of the body, such as the oral cavity, nasal passages, lungs, gut, skin, bladder, and vagina. We found that PVD and cardiovascular disease were associated with lower microbiota diversity in the gut (i.e., α-diversity), while supplemental vitamin use was associated with higher α-diversity.
Animal	NCIT_C14182	A non-human living organism that has membranous cell walls, requires oxygen and organic foods, and is capable of voluntary movement, as distinguished from a plant or mineral. --- Human is a subordinate concept of Animal but for CLEF we distinguish only between humans and any other animal.	link	Instances of "... *model*", such as "*rodent model*", should **NOT** be annotated as *animal* entity mentions.	Although approximately 30% mice are resilient to chronic social defeat stress (CSDS), the role of gut microbiota in this is unknown. We further demonstrated the role of Htr1a using AAV-shRNA to downregulate Htr1a in the mPFC of CUS mice. We compared the 16S ribosomal RNA (rRNA) gene sequences retrieved from fecal samples between control, CUMS-vulnerable, and CUMS-resilient mice.

Entity Category	URI	Definition	Reference URL	Notes	Examples
Statistical Technique	NCIT_C19044	A method of analyzing or representing statistical data; a procedure for calculating a statistic.	link	Instances of "*randomized controlled trials*" and "*cohort studies*" are too generic and should **NOT** be annotated.	Pearson's correlation analysis was used to evaluate the association between bacterial taxa and psychotic symptoms. Linear Discriminant Analysis (LDA) revealed Ruminococcaceae as a discriminative feature.
Biomedical Technique	NCIT_C15188	Research concerned with the application of biological and physiological principles to clinical medicine. --- This category also includes the *assay* category, defined as a planned process with the objective of producing information about a material entity (the evaluant) by examining it.	link	Instances of "*microbiota analysis*" should **NOT** be annotated as *biomedical technique* entity mentions.	The 16S rRNA amplicon sequencing method was performed to determine the fecal composition of fecal microbiota. The intestinal permeability biomarker zonulin was measured using enzyme-linked immunosorbent assays.
Bacteria	NCBITaxon_2	One of the three domains of life (the others being Eukarya and ARCHAEA), also called Eubacteria. They are unicellular prokaryotic microorganisms which generally possess rigid cell walls, multiply by cell division, and exhibit three principal forms: round or coccal, rodlike or bacillary, and spiral or spirochetal.	link	Microorganism that is unicellular and prokaryotic	Here we demonstrated that stress resistance in mice was associated with more abundant Lactobacillus and Akkermansia in the gut, but less abundant Bacteroides, Alloprevotella, Helicobacter, Lachnoclostridium, Blautia, Roseburia, Colidextibacter and Lachnospiraceae NK4A136. The abundance of Akkermansia, Megamonas, Prevotellaceae NK3B31 group, and butyrate-producing bacteria, Lachnospira, Subdoligranulum, Blautia, and Dialister, and acetate-producing bacteria,

Entity Category	URI	Definition	Reference URL	Notes	Examples
					Streptococcus, in the gut microbiota of the MDD group was lower than that in the control (C) group.
Chemical	CHEBI_59999	A chemical substance is a portion of matter of constant composition, composed of molecular entities of the same type or of different types.	link	The list of chemicals reported in https://pubchem.ncbi.nlm.nih.gov/ can be used as a reference for this label.	Significant difference was not detected in the expression of neurotransmitter receptor genes in the prefrontal cortex with the administration of sodium butyrate compared to that of the control group. Cryptotanshinone (CPT), a bioactive compound derived from the traditional Chinese herb Salvia miltiorrhiza, exhibits promising antidepressant properties.
Dietary Supplement	MESH_68019587	Products in capsule, tablet or liquid form that provide dietary ingredients, and that are intended to be taken by mouth to increase the intake of nutrients. Dietary supplements can include macronutrients, such as proteins, carbohydrates, and fats; and/or MICRONUTRIENTS, such as VITAMINS; MINERALS; and PHYTOCHEMICALS.	link	A *dietary supplement* is distinct from *food* in that it supplements the diet, providing additional nutrients or compounds, while *food* constitutes part of the diet itself.	A potential therapeutic strategy for maintaining a healthy life is to address stress-induced health problems with botanicals or dietary supplements such as polyphenols. Emerging data also suggests, particularly in rodents, that dietary interventions such as omega-3 fatty acids and pre- and pro-biotics may buffer against the effects of stress on the gut microbiome, but more research is needed.

Entity Category	URI	Definition	Reference URL	Notes	Examples
Disease, Disorder, or Finding (DDF)	NCIT_C7057	A condition that is relevant to human neoplasms and non-neoplastic disorders. This includes observations, test results, history and other concepts relevant to the characterization of human pathologic conditions.	link	Instances of "*... response*" (e.g., "*stress response*") and "*... mechanism*" (e.g., "*pathophysiological mechanisms*"), should **NOT** be annotated as *DDF* entity mentions.	Furthermore, alterations in the gut microbiota composition in humans have also been linked to a variety of neuropsychiatric conditions, including depression, autism and Parkinson's disease. Imbalances of this neurotransmitter are associated with neurological diseases, such as Alzheimer's and Parkinson's disease, and psychological disorders, including anxiety, depression, and stress. Cognitive impairment has been observed in patients with various psychiatric disorders, including schizophrenia, major depressive disorder (MDD), and bipolar disorder (BD).
Drug	CHEBI_23888	Any substance which when absorbed into a living organism may modify one or more of its functions. The term is generally accepted for a substance taken for a therapeutic purpose, but is also commonly used for abused substances.	link	The Drugbank database https://go.drugbank.com/ can be used as a reference for this label.	Additionally, a fluoxetine (FLU) has been used as a reference antidepressive drug. Accumulating evidence suggests that the N-methyl-D-aspartate receptor (NMDAR) antagonist ketamine produces rapid and sustained antidepressant effects… The N-methyl-D-aspartate receptor antagonist (R,S)-ketamine has attracted attention as a rapidly acting antidepressant.

Entity Category	URI	Definition	Reference URL	Notes	Examples
Food	NCIT_C62695	A group of solid, semi-solid, and liquid substances which are consumed by humans and animals. --- Although we acknowledge that *dietary supplement* could conceptually fit within this category, we are considering these two as distinct.	link	In general, beverages should be annotated under this label.	Fermented foods contain some of these compounds, which can affect human health and mood. The majority of food consumed during their stay included unpasteurised milk and dairy products.
Gene	SNOMEDCT_67261001	A functional unit of heredity which occupies a specific position on a particular chromosome and serves as the template for a product that contributes to a phenotype or a biological function.	link	In many cases, the proteins encoded by genes retain the same name as the genes themselves. In these instances, annotators must use the context of the abstract to determine whether the reference is to the gene or to the protein encoded by that gene.	A species of the Romboutsia genus was co-associated with the species of Ruminococcus gnavus in an internetwork through four genes: METTL8, ITGB2, OTULIN, and PROSER3, with a strict threshold ($p < 5 \times 10\text{-}4$). Human microbiota transplantation induced an emotionally impaired phenotype in mice and alterations in GABA-, proline-, and extracellular matrix-related prefrontal cortex genes. Soluble epoxide hydrolase (coded by the Ephx2 gene) plays an important role in inflammation, which has been implicated in stress-related depression.
Human	NCBITaxon_9606	Members of the species Homo sapiens. --- This category includes all mentions of humans, even if they are not directly informative about biomedical processes or	link	Instances of "*... population*", such as "*pediatric population*",	Bipolar disorder is rare among populations that have not adopted contemporary Western lifestyles, which supports the hypothesis that bipolar disorder results from a

Entity Category	URI	Definition	Reference URL	Notes	Examples
		discoveries. For instance, mentions such as "*psychiatrist*", "*clinicians*", "*medical personnel*", and similar terms should also be annotated under this entity label.		should be annotated as human entity mentions. On the other hand, instances of "*... model*", such as "*MDD model*", should **NOT** be annotated.	mismatch between Homo sapiens's evolutionary and current environments. In this systematic review and meta-analysis, we sought to examine the effects of probiotics supplementation on brain-derived neurotrophic factor (BDNF) in adults. Additionally, breast milk microbiota correlated more significantly with infants' SCFAs in the breastfeeding group than in the mixed feeding group.
Microbiome	OHMI_0000003	This term refers to the entire habitat, including the microorganisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes (i.e., genes), and the surrounding environmental conditions.	link	Although we acknowledge that a *microbiota* is not the same as a *microbiome*, at the current stage, we want to annotate both *microbiomes* and *microbiota* as *microbiome*.	The human gut microbiome is involved in a bi-directional communication pathway with the central nervous system (CNS), termed the microbiota-gut-brain axis. Several studies have shown that the gut microbiome is associated with FC, but these studies have produced inconsistent findings, with few reflecting the relationship between the gut microbiome and metabolites. Using the latest genome-wide association study (GWAS) summary data of the oral microbiome, polygenic risk scores (PRSs) of 285 salivary microbiomes and 309 tongue dorsum microbiomes were conducted.

4.2. General Entity Annotation Rules

➔ **Do not add new entity categories**

- The dataset supports exactly 13 predefined entity categories, and this set must remain unchanged. Use only the categories already provided.
- When assigning a category to an entity, always select it from the existing MetaTron list. If you type a category manually, remember that category names are case-sensitive: entering a category with a different casing than the one already present will create a new, unintended category in the MetaTron list.

➔ **Annotate all concept types**

- Annotate all text spans corresponding to the thirteen defined entity categories [2].
- Verify that pre-annotated entity boundaries are correct (see below), making any necessary adjustments.

➔ **Entity span and boundary rules**

- Annotate the complete text span that accurately describes the entity. The text span should start from the first character of the first word and end at the last character of the last word. Examples:
  - "increases cortisol levels";
  - "leads to disability";
  - "studies suggest that alteration in short-chain fatty acids".
- Annotate using full words only. DO NOT select only part of a word or a whole word along with a part of an adjacent word.
- A "word" is defined as a portion of text that is delimited by a whitespace on both the left and right. Words connected by symbols such as '-', '_', etc., should be considered as a single word. Examples:
  - "b-sitosteryl" is a single word;

- "genus_ruminococcaceae" is a single word;
- "short-chain fatty acids" is composed of three words: short-chain, fatty, and acids;
- "oligofructose-treated db/db mice" is composed of three words: oligofructose-treated, db/db, and mice.

- In some texts, markdown characters such as **, *<b>*, *<sub>*, etc., are included. Annotators should leave out these characters if they are annotating only the words between them. However, if they are annotating those words along with others outside the markdown tags, the markdown characters should be included. Examples:
  - "The presence of Ruminococcacae increases the risk...";
  - "An increase of depressive disorders has been noticed after the COVID pandemic".
- DO NOT include punctuation at the beginning or end of the mention. If punctuation is within the mention, it should be retained.
  - "... alteration in short-chain fatty acids …"
  - "... leads to depression, that can cumulate …"
- Ensure minimal context is preserved to maintain the correct meaning. For example, DO NOT omit suffixes or qualifiers if doing so changes the meaning or classification of the entity. Examples:
  - "Becker muscular dystrophy gene" should be annotated as a *gene* entity, not just "Becker muscular dystrophy" which would be a *disease*;
  - "dGK kinase deficiency" should be annotated as a *disease*, not just "dGK kinase" which would be a *gene*.
- Include relevant contextual modifiers needed to capture the full and precise meaning of the entity mentions. Adjectives should be included with the entity mention, while nouns used as modifiers should be annotated separately. Examples:

- "We recruited 54 subjects, including 27 patients with MDD";
- "The alpha diversity indices of MDD patients are …";
- "... were significantly enhanced in EC-12 supplemented mice".
- "Male mice fed on a diet supplemented with EC-12 showed…".

◆ Annotate a word only if it represents the intended entity in the given context. DO NOT annotate if the context alters the meaning of the term. Example:

- "Gut microbial changes derive from…" ;
- "Gut microbiota analysis are conducted to..";
- "... plays a crucial role in stress reactivity over the life span".

◆ Annotate **composite entities** as a single entity if they belong to the same category. However, if entities belong to the same category but appear as a sequence, annotate them separately. Examples:

- "SMADs 1, 5, and 8" should be annotated as a *gene*;
- "breast or ovarian cancer" should be annotated as a *disease*;
- "b-sitosteryl and stigmasteryl linoleates" should be annotated as a *chemical*;
- In "Cytochrome P-450 genes (CYP1A1, CYP2A6, CYP2D6, and CYP2E1)," label each of "Cytochrome P-450 genes," "CYP1A1," "CYP2A6," "CYP2D6," and "CYP2E1" as separate *gene* entities.

➔ **Overlapping and ambiguous entities**

◆ DO NOT annotate overlapping entities. An entity cannot include words that are already part of another entity mention.

- Although we acknowledge that allowing overlapping annotations of entity mentions would be the best approach to better capture the complexity and nuances of biomedical texts, at the current stage, we do not permit overlapping annotations. This decision has been made to simplify the annotation process and reduce the ambiguity that could arise during manual annotation.

- In "anti-mouse IL-6 receptor antibody" annotators should label the entire text span as a *chemical*, while they DO NOT have to annotate "mouse" as an *animal*.

- A text span can only be assigned to one entity category. If an entity could belong to multiple categories, use the context within the sentence to determine the appropriate category to be assigned.
  - In "... blockade of interleukin-6 receptor in the periphery …" "interleukin-6" might be labeled as a *gene*, but since it is followed by the word "receptor" we label the entire "interleukin-6 receptor" as a *chemical*.
- If an entity mention could be annotated in multiple ways, always annotate the longest version. Examples:
  - "chronic sleep disorder" must be annotated in full, rather than just "chronic sleep disorder";
  - "major depressive disorder" must be annotated in full, rather than just "major depressive disorder".
- When determining the span to annotate and deciding whether to keep the longest version, use the defined annotation tools to search if the longest version is recognized in a well-established knowledge source.
- When determining the span to annotate, use the existence of reference acronyms in the literature as an indicator to keep the longest version of the entity mention. Examples:
  - "major depressive disorder" is, in the literature, associated with the acronym MDD;
  - "amyotrophic lateral sclerosis" should be annotated in full, rather than just "amyotrophic lateral sclerosis" or "amyotrophic lateral sclerosis", since the acronym "ALS" is defined for the entire condition.

➔ **Special cases and abbreviations**

- DO NOT annotate terms that are identical to the entity labels. For example, the term “disease” by itself *should not* be annotated as a *disease* entity, while the term “Parkinson disease” should be annotated as a *disease* entity.
- Annotate both the abbreviation and its long form separately, if possible. Example:
  - In “Prostaglandin E2 (PGE2)” annotate both “Prostaglandin E2” and “PGE2” as separate *chemical* entities [2].
- If the boundary of an entity comprises both the full name name and its abbreviation, it *should* be annotated as a single entity. Example:
  - "Deoxyguanosine kinase (dGK) deficiency" should be annotated as a single *disease* entity [2].
- DO NOT annotate words that are morphological variations of terms that would be entities. Examples:
  - Do not annotate "hypertensive" as a *disease*, even though it is an adjective form of "hypertension."
  - Do not annotate "VLCAD deficient" as a *disease*, even though it refers to "VLCAD deficiency."

➔ **Non-Annotable Mentions**

- Certain terms may appear to fall within one of the defined categories but should not be annotated because they do not represent actual instances of those entities.
  - Instances of “*xxx axis*”, such as “*gut-brain axis*” or “*hypotalamic-pituitary-adrenal axis*”, even if they might be interpreted as *anatomical locations*, do not actually fit appropriately into any of the available entity labels and should be excluded from annotation.
  - Instances of "*xxx system*," such as "*immune system*" or "*endocrinal system*," should be excluded from annotation. These systems are organizations of varying numbers and types of organs, arranged to

perform complex functions for the body. Therefore, they do not represent a specific anatomical location but rather a set of anatomical locations linked together, and should not be annotated as *anatomical location* entity mentions.

- Instances of "*xxx models*", such as "*human models*", "*animal models*", "*rodent models*", etc.. should not be annotated. Although they contain terms like "*human*" or "*animal*", these refer to experimental models rather than actual instances of human or animal entities.
- Instances of "*xxx response*" and "*xxx mechanism*", such as "*stress response*" or "*pathophysiological mechanism*", describe biological or physiological processes rather than specific *DDF* entity mentions and, therefore, should be excluded from annotation.
- Instances of "*microbiota analysis*" are too generic and do not refer to specific biomedical techniques. Therefore, such mentions should not be annotated.
- Instances of "*randomized controlled trials*" and "*cohort studies*" are more aligned with research methods rather than specific biomedical or statistical techniques. Therefore, such mentions should not be annotated.
- Instances of "*effects of xxx*" or "*xxx therapy*" refer to therapeutic aspects or treatment outcomes, not to actual DDFs. Therefore, these mentions should not be annotated as DDF entities.
- Instances of "*xxx symptoms*" or *"xxx pathogenesis*" describe clinical manifestations or causal mechanisms rather than specific DDFs. These should not be annotated as DDF entities.
- Instances of "*xxx destruction*" (e.g., neuronal destruction) refer to biological processes, not to specific DDFs. Therefore, such mentions should not be annotated.

➔ **Use of full text and tools**

- Annotators can access the full text and use various tools detailed in the "Annotation Tools" section to clarify entity boundaries and labels [2]. In particular, we recommend the use of:
  - Ontology Lookup Service (OLS)
  - Ontobee
  - UMLS Metathesaurus Browser
- If these tools are not sufficient to clarify their doubts, annotators are free to search the internet for more information. However, they must pay careful attention to the reliability of the websites being consulted, prioritizing reputable and authoritative sources.

4.3. Entity Linking Rules

➔ **Priority order of reference resources**

- Link each identified entity mention to a concept following the defined priority order for its entity category, as described in the table below.
- If a concept is not available in the reference resources specified for its entity category but can be found in another reputable biomedical resource, you may still include it. However, always prioritize concepts from the higher-priority resources whenever possible.

Entity Category	Linked Vocabularies (in priority order)
Anatomical Location	1. NCIT 2. UMLS 3. GBIE
Animal	1. NCIT 2. NCBITaxon 3. UMLS 4. GBIE
Bacteria	1. NCIT 2. NCBITaxon 3. MESH 4. OMIT 5. UMLS 6. GBIE
Biomedical Technique	1. NCIT 2. OMIT 3. NCBITaxon 4. UMLS 5. GBIE
Chemical	1. NCIT 2. CHEBI 3. OMIT 4. UMLS 5. GBIE
Dietary Supplement	1. NCIT 2. CHEBI 3. NCBITaxon 4. OMIT 5. MESH 6. UMLS 7. GBIE
DDF	1. NCIT 2. OMIT 3. NCBITaxon 4. UMLS 5. GBIE
Drug	1. NCIT 2. CHEBI 3. OMIT 4. NCBITaxon 5. UMLS 6. GBIE
Food	1. NCIT 2. UMLS 3. GBIE
Gene	1. NCIT 2. OMIT 3. CHEBI 4. UMLS 5. GBIE
Human	1. NCIT 2. MESH 3. UMLS 3. GBIE
Microbiome	1. NCIT 2. NCBITaxon 3. OHMI 4. UMLS 5. GBIE
Statistical Technique	1. NCIT 2. STATO 3. SWO 4. UMLS 5. GBIE

➔ **Concept selection**

- Always link to the most specific concept available, the one that most accurately and precisely describes the entity mention.
- Each entity category includes a "*most generic*" concept that represents its broadest possible meaning. You should use this concept only if no more specific concept is available for the given entity mention. Linking to one of these generic concepts when a more specific one exists is considered an annotation error.

Category	Concept Name	Ontology ID
Animal	Animal	NCIT:C14182
Bacteria	Bacteria	NCBITaxon:2
Biomedical Technique	Biomedical Computing	NCIT:C19247
Chemical	Chemical	NCIT:C48807
Dietary Supplement	Dietary Supplement	NCIT:C1505
Disease or Disorder	Disease or Disorder	NCIT:C2991
Drug	Drug	CHEBI:23888
Finding	Finding	NCIT:C3367
Food	Food	NCIT:C62695
Gene	Gene	NCIT:C16612
Human	Human	NCIT:C14225
Metabolite	Metabolite	CHEBI:25212
Microbiome	Microbiome	OHMI:0000003
Neurotransmitter	Neurotransmitters	UMLS:C0027908
Statistical Technique	Statistical Technique	UMLS:C1710191

- Documents often contain semantic variations of the same concept (e.g., "*Interleukin-10*" vs "*IL-10*"). Before adding a new concept, always check

whether an equivalent form, such as an acronym, long form, or alternative spelling, already exists in the MetaTron list. This helps prevent duplicated concepts referring to the same underlying entity.

- Linking entity mentions to concepts greatly facilitates systematic correction during post-processing. Therefore, if you are uncertain about whether a mention should be included or which concept to choose, annotate it and link it to the best candidate concept. After the annotation phases, we will manually review all used concepts and apply the necessary corrections or removals.

➔ **Missing and unavailable concepts handling**

- When a suitable concept cannot be found, distinguish between three cases:
  - 1. The concept exists in a reference resource but is missing from our current MetaTron concept list → add the concept to MetaTron under the appropriate entity category. CHECK THIS GUIDE TO SEE HOW TO CORRECTLY ADD NEW CONCEPTS.
  - 2. The concept does not exist in any reference resource → link the mention to the *most generic concept* for its entity category.
  - 3. The concept exists and is included in the MetaTron concept list, but is associated with an incorrect entity category → assign the correct concept and select one of the available categories for that concept.
- When a mention is linked to a concept under one of the three cases above, annotators are required to note down in the shared google sheet:
  - The PMID of the document,
  - The text span and entity label,
  - Any other relevant contextual information
- This information will be used in the manual post-processing stage to:
  - Review concepts added to the MetaTron list.
  - Create new entries in the GutBrainIE custom ontology.
  - Update concept-category associations where necessary.

- The shared google sheet contains two sheets "*Added Concepts*" and "*Custom Concepts*":
  - In "*Added Concepts*" you should record all concepts added to the MetaTron list taken from biomedical resources.
    - This is no longer needed and will be done in post-processing after completing the manual curation phases. It is still fundamental you record in the sheet concepts defined for our custom ontology.
  - In "*Custom Concepts*" you should include all new concepts that need to be defined and included in our custom ontology.
- The shared google sheet also contains a "*Duplicated Concepts*" sheet. This should be used to report concepts with different URIs but identical meaning, i.e., duplicated that should be merged into a single concept during post-processing.

➔ **Use of additional reference resources**

- If a concept is not available in the defined reference resources but is available in another reliable biomedical resource, you might include and use that concept for linking. These cases will be reviewed in post-processing to minimize the number of external resources employed.

➔ **Use of the DAIsO ontology for ALS-related mentions**

- For concepts related to Amyotrophic Lateral Sclerosis (ALS), you might consider also the DAIsO ontology, which contains several relevant concepts, particularly for DDF mentions. All DAIsO ontology concepts are listed in the shared spreadsheet at sheet "DAIsO ontology" (red color).

➔ **Use of the "Diagnostic and Statistical Manual of Mental Disorders (DSM)" for Custom Concepts**

- For DDF mentions related to mental disoders that are not included in any biomedical resource you might refer to the "Diagnostic and Statistical Manual

of Mental Disorders (DSM)" to define custom concepts. The manual can be accessed at this Google Drive link.

➔ **Composite mentions**

- When a mention includes multiple entities of the same category (e.g., "<DDF> and <DDF> diseases"), link it to the most specific concept that includes both
  - "... neurodegenerative disorders such as Alzheimer's and Parkinson's diseases." → Neurodegenerative Disorders (UMLS:C0524851)
  - "... cardiovascular diseases (ischemic and non-ischemic heart diseases) ..." → Cardiovascular Diseases (UMLS:C0007222)

➔ **Cancer-related mentions**

- Treat *cancer* and *carcinoma* as distinct concepts.
- When formulated as "<anatomical location> cancer" (e.g., breast cancer, lung cancer), annotate as a single mention and link to the most specific concept available:
  - Breast cancer → Breast Cancer (MONDO:0007254)
  - Lung cancer → Lung Cancer (MONDO:0008903)
- When formulated as "cancer of the <anatomical location>", annotate two mentions:
  - Cancer of the lung → Cancer (NCBITaxon:6754); Lung (NCIT:C12468)
  - Annotate a "strike" relation from DDF (cancer) to anatomical location (lung).

➔ **Mice-related mentions**

- When the mention specifies gender or genetic variation of the mouse, link to the most specific concept available. If none exists, use Mus musculus (NCIT:C45247)
  - "... from young and estropausal female mice ..." → Female Mus musculus [HERO]

- "... increased serum indole-3-propionic acid levels in male mice ..." → Male Mus musculus [HERO]
- "... 5 different mice breeds: Balb/c, Orient C57BL/6N ..." → BALB/c Mouse [NCIT]; C57BL/6N Mouse [NCIT]

➔ **"Altered gut microbiota" mentions**

- Mentions such as "altered gut microbiota" or variations (e.g., "gut microbiota alteration", "disrupted gut microbiota") should be linked to "Dysbiosis" [OMIT:0028544]

➔ **"Stool collection" mentions**

- Mentions referring to "stool collection" or "stool collection techniques" should be linked to a specific concept in a reference resource. If no specific concept exists, link to the generic "Stool collection" [UMLS:C5849121]
- When annotating mentions referring to *stool collection*, distinguish between procedure and sample references:
  - If referring to the collection process or technique, label as biomedical technique and link to "Stool Collection" (UMLS:C5849121)
  - If referring to the collected samples, label as chemical and link to "Stool Specimen" (UMLS:C1550661)

➔ **"xxx-patients" mentions**

- Mentions of "xxx patients" (e.g., "Parkinson's patients, ALS patients") should be linked to the generic concept "Patients" (MESH:68010361)

➔ **"<Microbiome/Microbiota>" mentions**

- "<Anatomical location> + <microbiome/microbiota>" mentions, such as "gut microbiome/microbiota" or "oral microbiome/microbiota" should be linked to the specific concept representing that anatomical location's microbiome if available in reference resources. If none exists, define a new concept to be instantiated in the GutBrainIE custom ontology.

- the role and mechanism of oral microbiota in the development of multiple sclerosis are still elusive → Oral microbiome [NCIT:C125208]
- Gut microbiota regulate Alzheimer's disease pathologies and cognitive disorders via PUFA-associated neuroinflammation → human gut microbiome [OHMI:0000020]

- When annotating mentions of microbiome/microbiota, if the text explicitly or implicitly refers to the human microbiome, link to the specific human microbiome/microbiota concept. Otherwise, link to the generic "Microbiota" [OHMI:0000003]

➔ **"<DDF>-like behavior" and "<DDF>-related behavior" mentions**

- Mentions of "<DDF>-like behavior" and "<DDF>-related behavior" should always be linked to a specific concept. If no concept is available from reference resources, define a new one for our custom ontology.
  - "... decreases in anxiety-like behaviours" → Anxiety-like Behavior (NBO:0000092)
  - "... no alterations in depression-like behaviours ..." → Depression-like Behavior (MP:0003360)
- For a given DDF, instances of "<DDF>-like behavior" and "<DDF>-related behavior" should be linked to the same concept.

➔ **"xxx-induced inflammation" mentions**

- Mentions of "xxx-induced inflammation" should always be linked to the general concept "Inflammation" (NCIT:C3137)
  - "... protection against bacterial DNA-induced inflammation."
  - "... effects of obesity and diabetes-induced inflammation on the BBB."
  - "... decreased LPS-induced inflammation in BV2 cells."

➔ **Mentions referring to groups of people**

- Mentions referring to human populations should be linked to “General Population” (UMLS:C0681558) unless the text explicitly refers to a specific subgroup.
  - “... as a mainly elderly population suffering from ...” → Elderly (population group) [UMLS:C0681558]
- Mentions such as “human subjects” should be linked to “General Population” (UMLS:C0681558)
  - “... data derived from human subjects ...” → General Population [UMLS:C0681558]
- Mentions of “placebo group” should be linked to “Control Group” [NCIT:C28294].

5. Relation Annotation Guidelines

5.1. Relation Labels

Head Entity	Tail Entity	Predicate
Anatomical Location	Human / Animal	**located in**
Bacteria	Bacteria / Chemical / Drug	**interact**
Bacteria	Disease, Disorder, or Finding	**Influence**
Bacteria	Gene	**change expression**
Bacteria	Human / Animal	**located in**
Bacteria	Microbiome	**part of**
Chemical	Anatomical Loc. / Human / Animal	**located in**
Chemical	Chemical	**interact / part of**
Chemical	Microbiome	**impact / produced by**
Chemical / Dietary Supp. / Drug / Food	Bacteria / Microbiome	**impact**
Chemical / Dietary Supp. / Food	Disease, Disorder, or Finding	**influence**
Chemical / Dietary Supp. / Drug / Food	Gene	**change expression**
Chemical / Dietary Supp. / Drug / Food	Human / Animal	**administered**
Disease, Disorder, or Finding	Anatomical Location	**strike**
Disease, Disorder, or Finding	Bacteria / Microbiome	**change abundance**
Disease, Disorder, or Finding	Chemical	**interact**
Disease, Disorder, or Finding	Disease, Disorder, or Finding	**affect / is a**
Disease, Disorder, or Finding	Human / Animal	**target**
Drug	Chemical / Drug	**interact**
Drug	Disease, Disorder, or Finding	**change effect**
Human / Animal / Microbiome	Biomedical Technique	**used by**
Microbiome	Anatomical Loc. / Human / Animal	**located in**
Microbiome	Gene	**change expression**
Microbiome	Disease, Disorder, or Finding	**is linked to**
Microbiome	Microbiome	**compared to**
*	*	**associated with**

5.2. General Relation Annotation Rules

➔ **Relation types to annotate**

- Annotate relations between entities only if they match the defined set of relations provided for the *GutBrainIE* dataset.
- In certain cases, a relation may exist between two entity mentions that cannot be captured by any of the defined valid relations. In these instances, annotators may use the label "*associated with*". Examples:
    - In "The concentration of Bifidobacterium in the gut is studied using droplet digital PCR on fecal samples." there are two potential relations: one between "bifidobacterium" (*bacteria*) and "droplet digital PCR" (*biomedical technique*), and another between "bifidobacterium" (*bacteria*) and "gut" (*anatomical location*). Since no valid relations are defined between *bacteria* and *anatomical location*, or between *bacteria* and *biomedical technique*, these relations should be annotated using the label *in associated with*.
    - In "abnormal concentrations of firmicutes have been proven to have dramatic negative effects on the gut microbiota", it is clear that a relationship exists between "firmicutes" (*bacteria*) and "gut microbiota" (*microbiome*). However, the only valid relation defined between these entity types is *part of*, which does not hold in this context. Therefore, no relation should be annotated.
- The head of a relation does not always precede the tail in the text; it may also come after the tail. Annotators should ensure that the correct entities are linked regardless of their order in the text. Examples:
    - In "Depression has been historically treated through antidepressant medications. Nowadays, probiotics supplementation is being used to…" two *change effect* relations should be annotated, one from

"antidepressant medications" (*drug*) and "depression" (*disease*), and the other from "probiotics supplementation" (*dietary supplement*) to "depression" (*disease*).

- Annotate relations that are directly and explicitly mentioned in the text. Example:
  - In "Firmicutes influences predisposition to major depressive disorder" an *influences* relation between "Firmicutes" (*bacteria)* and "major depressive disorder" (*disease*) should be annotated
- Annotate relations implied by the text, even if the specific term describing the relation is not used. Example:
  - In "Firmicutes impact inflammation in the gut" an *influences* relation should be annotated between "Firmicutes" (*bacteria)* and "inflammation" (*disease*), as it is implied by the verb "impact".
- Annotate relations that can be inferred from the context. Relations inferred by reasoning about the text are valid as long as there is clear contextual evidence in the text and no personal, previous, or external knowledge is used for that inference. Example:
  - In "Firmicutes change the expression of gene ARID1B and, therefore, affect predisposition to Autism spectrum disorder (ASD) in young patients" it can be inferred that "Firmicutes" (*bacteria*) is related to "Autism spectrum disorder (ASD)" (*disease*) since it plays a role in affecting the "gene ARID1B" (*gene*) related to its predisposition. Therefore, two relations should be annotated In this portion of text: the explicit one *change expression* between "Firmicutes" (*bacteria*) and "gene ARID1B" (*gene*), and the inferred one *influence* between "Firmicutes" (*bacteria*) and "Autism spectrum disorder (ASD)" (*disease*).

- In “Bacteroides fragilis produces gamma-aminobutyric acid (GABA). GABA has been found to reduce anxiety in mice.” it can be inferred, although not explicitly stated, that “Bacteroides fragilis” (*bacteria*) has a role in reducing “anxiety” (*disease*) through its production of “gamma-aminobutyric acid (GABA)” (*chemical*). Therefore, an *influence* relation can be inferred between "Bacteroides fragilis" (*bacteria*) and "anxiety" (*disease*).

➔ **Contextual requirements for annotating relations**

- Co-occurrence is not required. The entities involved in a relation do not need to co-occur in the same sentence. Relations can be annotated even if the entities are in different parts of the abstract, as long as there is sufficient contextual information to support the relation. Example:
  - In “The gut microbiome is known to produce short-chain fatty acids. [...] The latter populate the intestinal barrier and play a crucial role in maintaining its integrity.” an inferred *produced by* relation can be annotated between “short-chain fatty acids” (*chemical*) and “gut microbiome” (*microbiome*) from the first sentence, and a *located in* relation between “short-chain fatty acids” (*chemical*) and “intestinal barrier” (*anatomical location*) can be annotated although they do not co-occur in the same sentence.
- Avoid overgeneralization, namely, only annotate relations between specific entity instances if there is explicit, implicit, or inferred evidence of their connection within the abstract. Do not assume that all occurrences of the same entity pair have a relation unless it is explicitly or implicitly described. Example:
  - If one mention of “gut microbiome” says “The gut microbiome is associated with reduced symptoms of Parkinson's disease” and another simply states “Parkinson's disease is prevalent among the

elderly" only the former should be used to annotate the relation *is linked to*.

- If there is a relation from a long version of an entity to another entity, the same relation should be annotated from the short/acronym version of the entity to the same target entity. Example:
  - In "Selective serotonin reuptake inhibitors (SSRIs) have an important role in the pathogenesis of depression." two relations *change effect* should be annotated, one from "Selective serotonin reuptake inhibitors" (*drug*) to "depression" (*disease*), and a second one from "SSRIs" (*drug*) to "depression" (*disease*).

➔ **Annotation context and scope**

- Annotators should determine if a relation between two entities holds by only considering the information presented in the abstract. Annotators are not allowed to access the full text of the article nor to use any external sources of information. If the abstract is not clear about the relationship, DO NOT annotate it.

➔ **Doubtful cases**

- If annotators are uncertain whether a relation exists between two entities, they should annotate conservatively. Only annotate when there is sufficient evidence to support a direct, implied, or inferred relationship.

6. Annotation Tools

To assist annotators in performing their tasks more effectively, the following tools are recommended. These tools fall into three categories:

6.1) Automatic annotation tools for entities

6.2) Ontology and database search tools

6.3) Category-specific search tools

The tools within each category are listed in descending order of reliability and importance. We highly recommend prioritizing the highest-ranked tools.

The table reported below summarizes the tools recommended in the following sections

Tool	Category	NER	NEN	RE	Entity Types
PubTator3	Automatic Annotation	☑	☑	☑	gene, disease, chemical, variant, species, and cell line
BERN2	Automatic Annotation	☑	☑	☐	gene, disease, chemical, species, mutation, cell line, cell type, DNA, and RNA
OnTheFly2.0	Automatic Annotation	☑	☑	☐	protein, chemical, organism, environment, tissue, disease, and gene
MedCat	Automatic Annotation	☐	☑	☐	Links concepts to UMLS and SNOMED CT
Bio-NLP	Automatic Annotation	☑	☑	☐	disease, chemical, and gene
OntoBee	Ontologies/Databases Search	☐	☐	☐	all the *GutBrainIE* entity labels
OLS	Ontologies/Databases Search	☐	☐	☐	all the *GutBrainIE* entity labels
Alliance of Genome Resources	Ontologies/Databases Search	☐	☐	☐	organism and gene
ChEBI	Category-Specific Search	☐	☐	☐	chemical
AmiGO2	Category-Specific Search	☐	☐	☐	gene
BacDive	Category-Specific Search	☐	☐	☐	bacteria
NCBI Taxonomy	Category-Specific Search	☐	☐	☐	organism
GeneCards	Category-Specific Search	☐	☐	☐	gene
DrugBank	Category-Specific Search	☐	☐	☐	drug
PubChem	Category-Specific Search	☐	☐	☐	chemical
Panther	Category-Specific Search	☐	☐	☐	gene

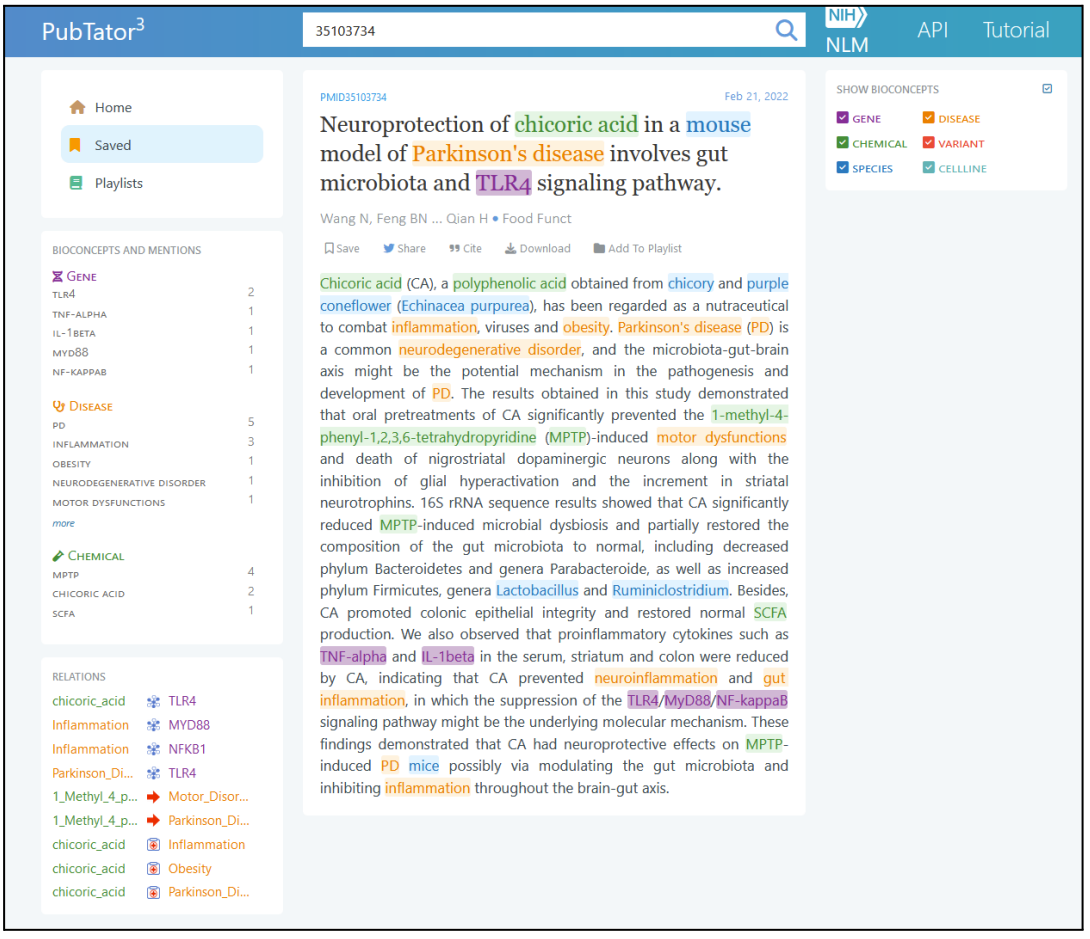
6.1. Automatic Annotation Tools for Entities

These tools provide automated annotation of entities for different biomedical categories, producing visual outputs that can assist annotators. These tools can be used in conjunction with the pre-annotations uploaded on *MetaTron*.

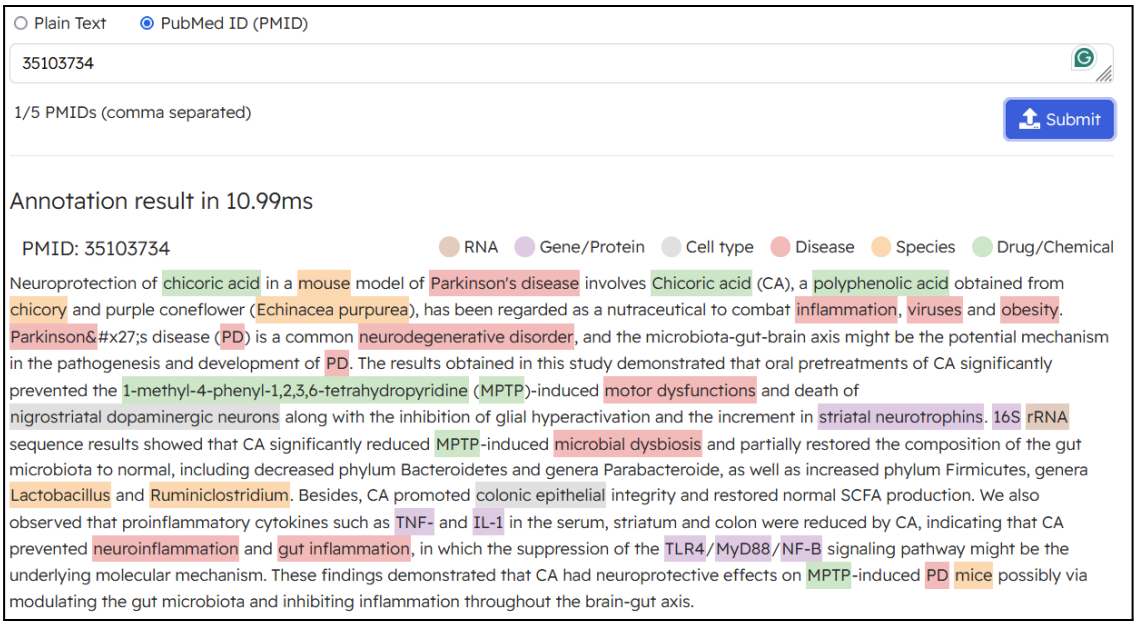
We suggest that annotators keep these automated annotations alongside *MetaTron*, using them as a reference to cross-check, verify, and complement the pre-annotations.

➔ **PubTator3**: a well-regarded tool for automatically annotating biomedical abstracts for NER and NEN, considering the following entities: gene, disease, chemical, variant, species, and cell line.

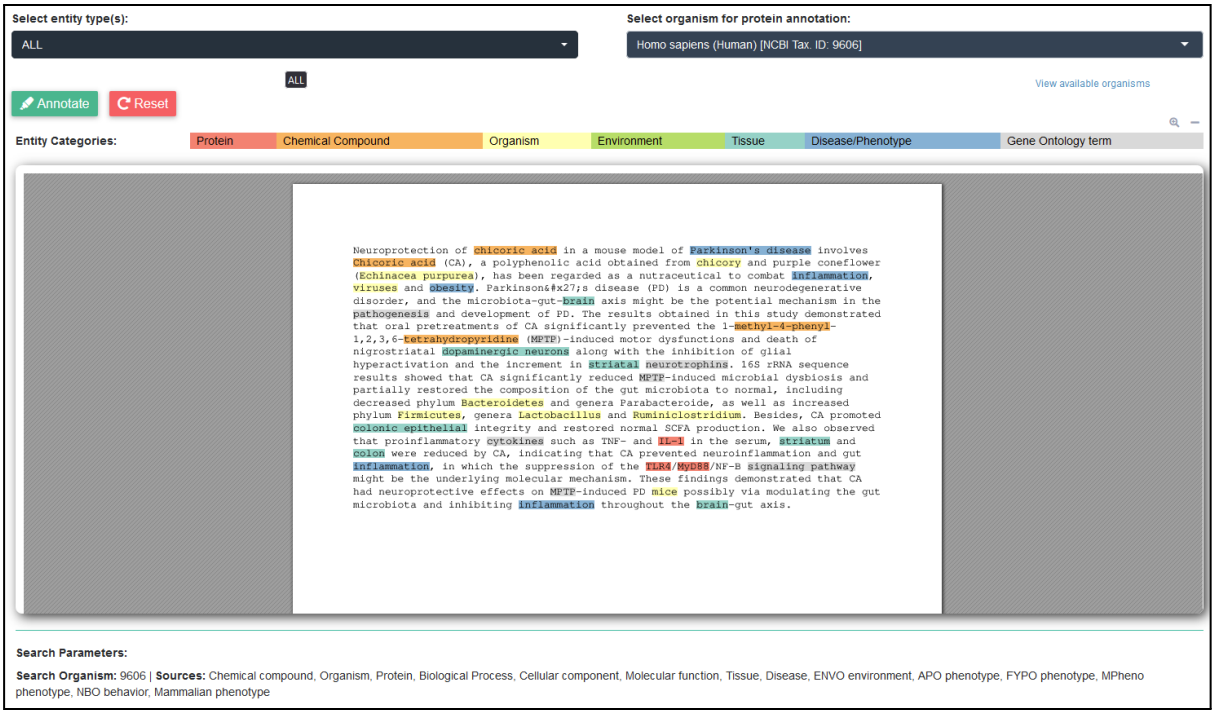
It requires the PMID of the articles in input, and it returns the abstract annotated with easy-to-interpret visual outputs. Additionally, PubTator can automatically annotate certain types of relations, including gene-disease associations, chemical-disease interactions, and gene-chemical interactions.



➔ **BERN2**: a tool that supports automated biomedical NER and NEN for: gene, disease, chemical, species, mutation, cell line, cell type, DNA, and RNA. It provides easy-to-interpret visual outputs. It takes in input either plain text or the article PMID.



➔ [OnTheFly2.0]: a tool that provides visual NER and NEN for: protein, chemical, organism, environment, tissue, disease, and gene. It requires in input either a PDF or plain text.



➔ [MedCAT]: a medical concept annotation tool that links entities to medical ontologies such as UMLS and SNOMED CT, allowing to perform NER considering the type and NEN considering the identifier.

Neuroprotection of chicoric acid in a mouse model of Parkinson's disease involves Chicoric acid (CA), a polyphenolic acid obtained from chicory and purple coneflower (Echinacea purpurea), has been regarded as a nutraceutical to combat inflammation , viruses and obesity . Parkinson's disease (PD) is a common neurodegenerative disorder , and the microbiota-gut- brain axis might be the potential mechanism in the pathogenesis and development of PD. The results obtained in this study demonstrated that oral pretreatments of CA significantly prevented the 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)-induced motor dysfunctions and death of nigrostriatal dopaminergic neurons along with the inhibition of glial hyperactivation and the increment in striatal neurotrophins. 16S rRNA sequence results showed that CA significantly reduced MPTP-induced microbial dysbiosis and partially restored the composition of the gut microbiota to normal , including decreased phylum Bacteroidetes and genera Parabacteroide, as well as increased phylum Firmicutes, genera Lactobacillus and Ruminiclostridium. Besides, CA promoted colonic epithelial integrity and restored normal SCFA production. We also observed that proinflammatory cytokines such as TNF- and IL-1 in the serum, striatum and colon were reduced by CA, indicating that CA prevented neuroinflammation and gut inflammation , in which the suppression of the TLR4/MyD88/NF-B signaling pathway might be the underlying molecular mechanism. These findings demonstrated that CA had neuroprotective effects on MPTP-induced PD mice possibly via modulating the gut microbiota and inhibiting inflammation throughout the brain -gut axis .

Pretty Name	Butanedioic Acid 2 3 Bis 2E 3 3 4 Dihydroxyphenyl 1 Oxo 2 Propen 1 Yl Oxy 2R 3R
Identifier	C0391039
Type	Pharmacologic Substance
Confidence Score	1
Start Index	19
End Index	32
ICD-10 Code	-
id	0
Status	Affirmed

➔ [Bio-NLP]: a tool to perform NER and NEN for disease, chemical, and genetic entities from plain text.

Results

Neuroprotection of chicoric acid CHEMICAL in a mouse model of Parkinson's disease DISEASE involves Chicoric acid CHEMICAL (CA), a polyphenolic acid CHEMICAL obtained from chicory and purple coneflower (Echinacea purpurea), has been regarded as a nutraceutical to combat inflammation DISEASE , viruses DISEASE and obesity DISEASE . Parkinson's disease (PD DISEASE) is a common neurodegenerative disorder DISEASE , and the microbiota-gut- brain axis might be the potential mechanism in the pathogenesis and development of PD DISEASE . The results obtained in this study demonstrated that oral pretreatments of CA significantly prevented the 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine CHEMICAL (MPTP)-induced motor dysfunctions and death DISEASE of nigrostriatal dopaminergic neurons along with the inhibition of glial hyperactivation and the increment in striatal neurotrophins GENETIC . 16S rRNA sequence results showed that CA significantly reduced MPTP CHEMICAL -induced microbial dysbiosis DISEASE and partially restored the composition of the gut microbiota to normal , including decreased phylum Bacteroidetes and genera Parabacteroide, as well as increased phylum Firmicutes, genera Lactobacillus and Ruminiclostridium. Besides, CA promoted colonic epithelial integrity and restored normal SCFA production. We also observed that proinflammatory cytokines such as TNF- and IL-1 GENETIC in the serum, striatum and colon were reduced by CA, indicating that CA prevented neuroinflammation DISEASE and gut inflammation DISEASE , in which the suppression of the TLR4 GENETIC / MyD88 GENETIC / NF-B GENETIC signaling pathway might be the underlying molecular mechanism. These findings demonstrated that CA had neuroprotective effects on MPTP CHEMICAL -induced PD DISEASE mice possibly via modulating the gut microbiota and inhibiting inflammation DISEASE throughout the brain -gut axis .

Normalized found terms

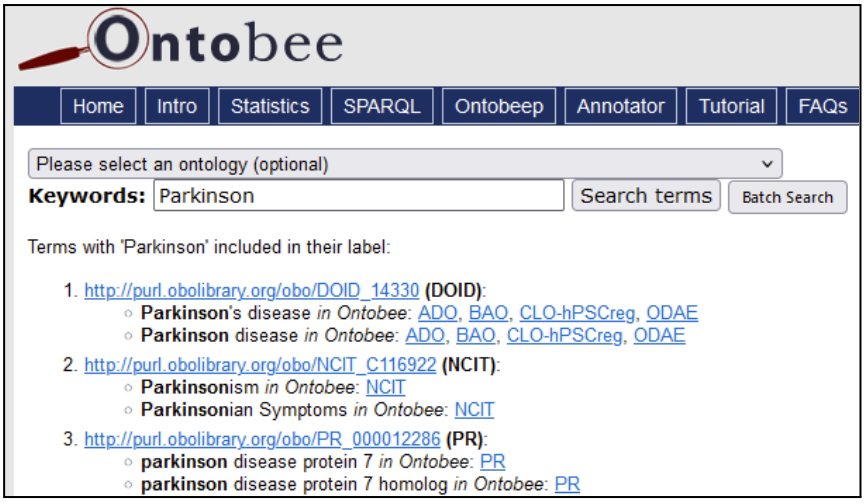
Diseases

Term	Found Term	Mesh ID	CUI	ICD10	Other Ids	Semantic type
Parkinson's disease	Parkinson's disease	-	C0865474, C0865475	G20	-	Disease or Syndrome
inflammation	Neurogenic Inflammation	D020078	-	-	-	Nervous system disease | Pathology (process)
viruses	Viruses	-	-	-	NCBITaxon: 10239	-
obesity	Pediatric Obesity	D063766	-	-	-	Nutrition disorder | Signs and symptoms

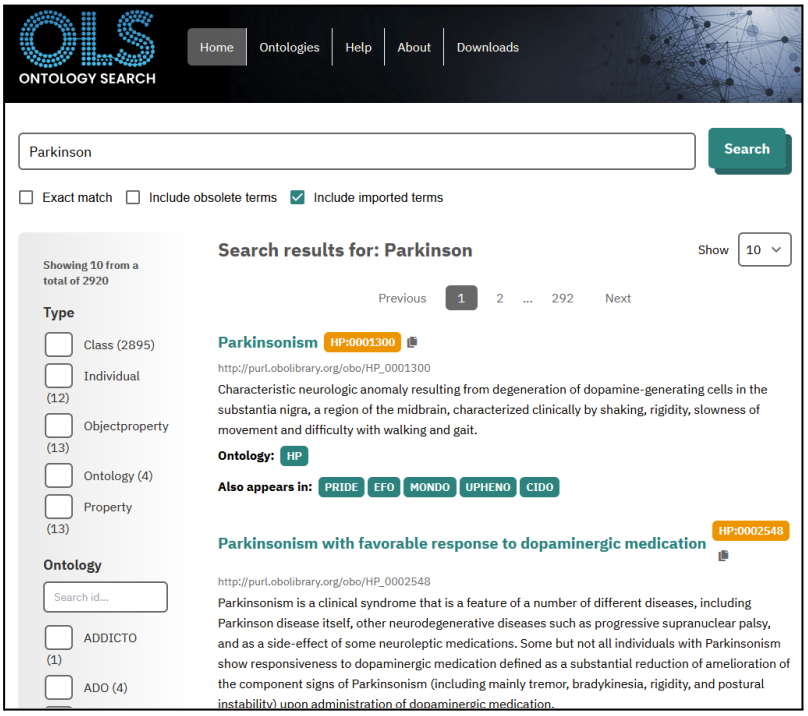
6.2. Ontology and Database Search Tools

These tools allow for the search of terms from multiple biomedical ontologies and databases, useful for validating and verifying entities.

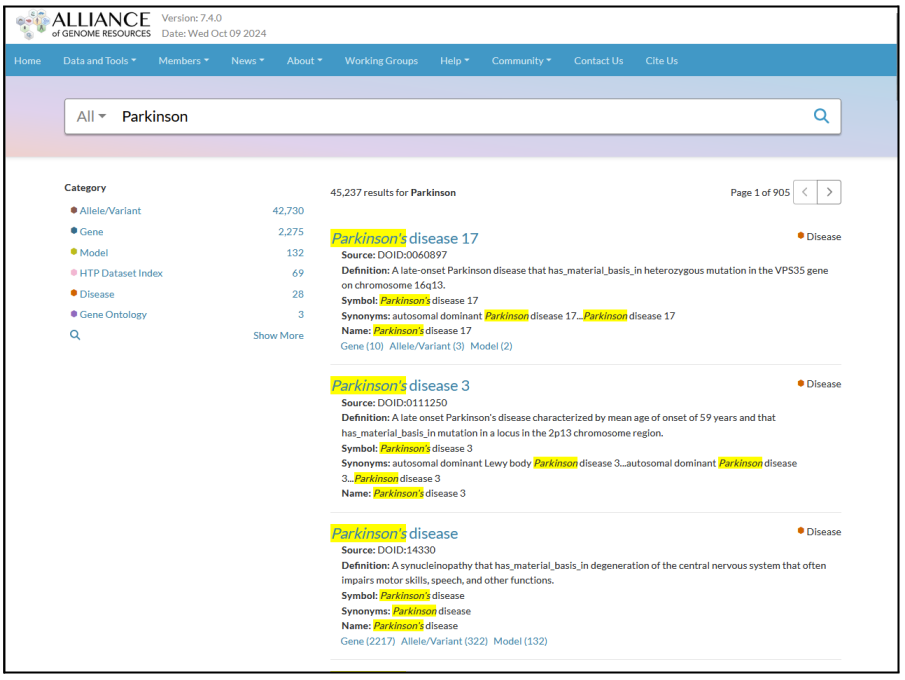
➔ **OntoBee**: a platform that enables searching across more than 250 biomedical ontologies, covering all the entities of interest for *GutBrainIE*.



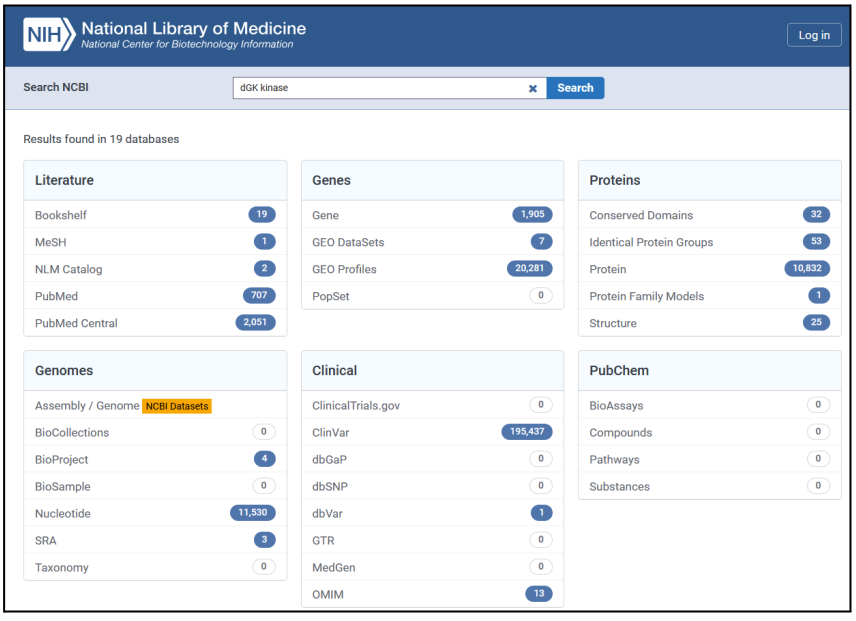
➔ **OLS Ontology Search**: similar to OntoBee, it allows searching across more than 250 biomedical ontologies, covering all the entities of interest for GutBrainIE.



➔ **Alliance of Genome Resources**: The Alliance of Genome Resources is a consortium of seven Model Organism Databases (MODs) and the Gene Ontology (GO) Consortium



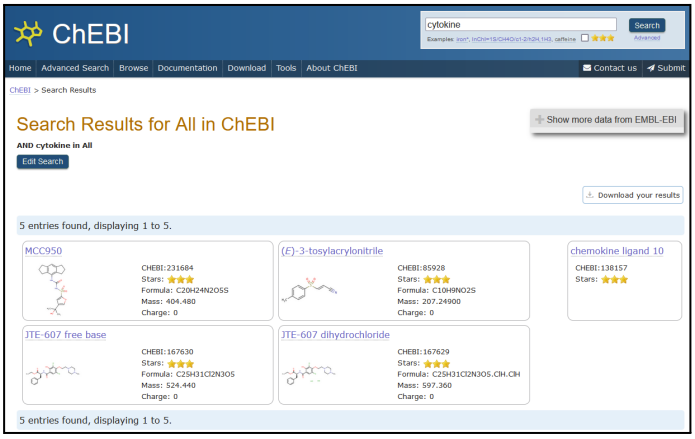
➔ **NCBI (National Center for Biotechnology Information)**: NCBI provides comprehensive access to a wide array of databases, including gene, protein, and taxonomy resources.



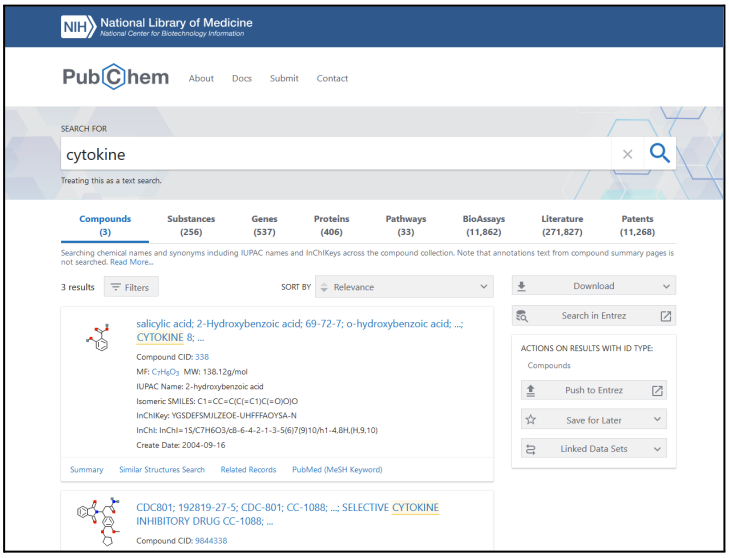
6.3. Category-Specific Search Tools

These tools are specialized for searching entities in specific categories, such as chemicals, genes, and diseases.

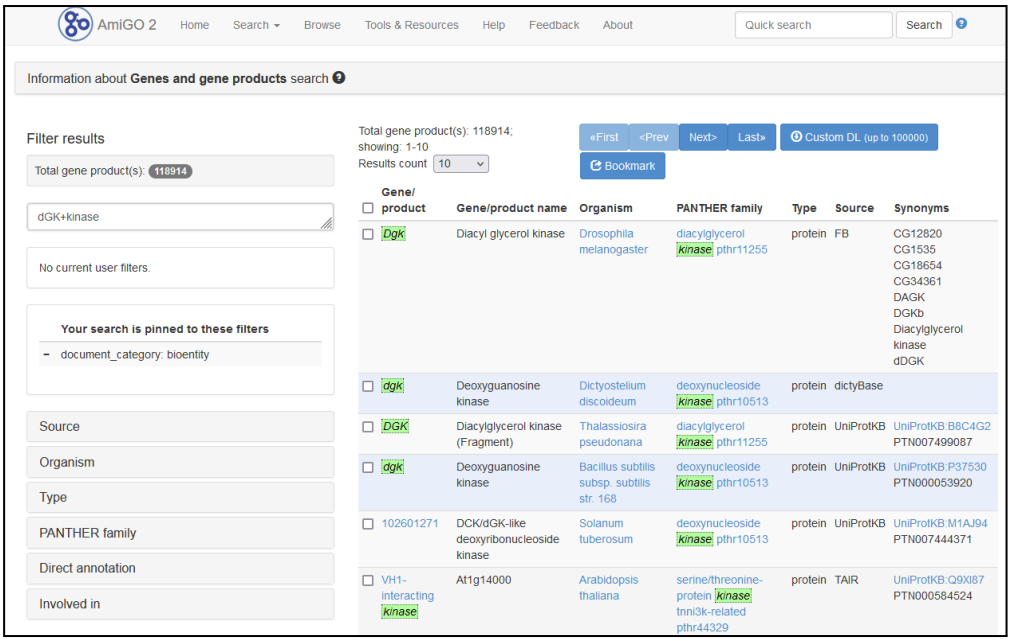
- **ChEBI**: The Chemical Entities of Biological Interest (ChEBI) database is a high-quality source for finding chemical information, including chemical properties and biological roles. It is the primary recommendation for annotating chemical-related entities.



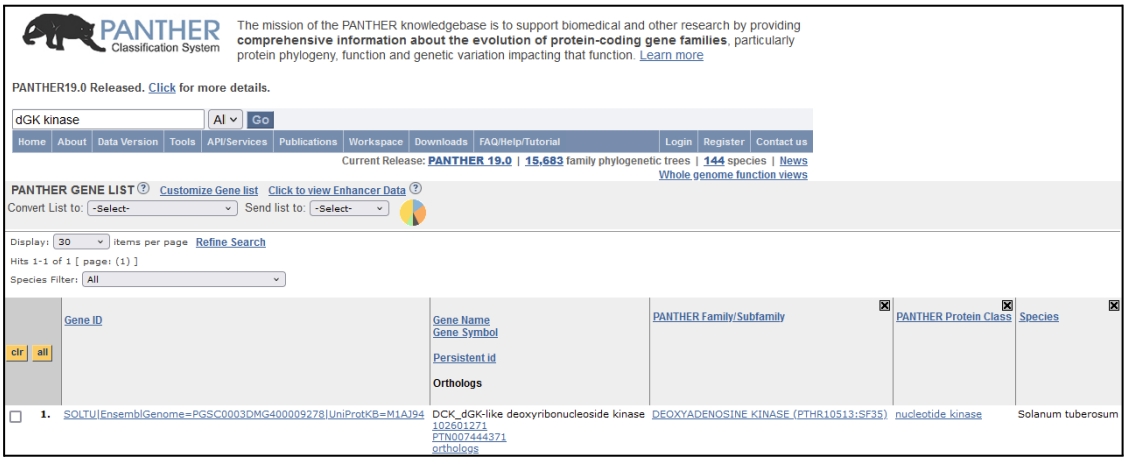
- **PubChem**: PubChem contains extensive information on chemical substances, including their properties and activities, making it ideal for verifying chemical annotations.



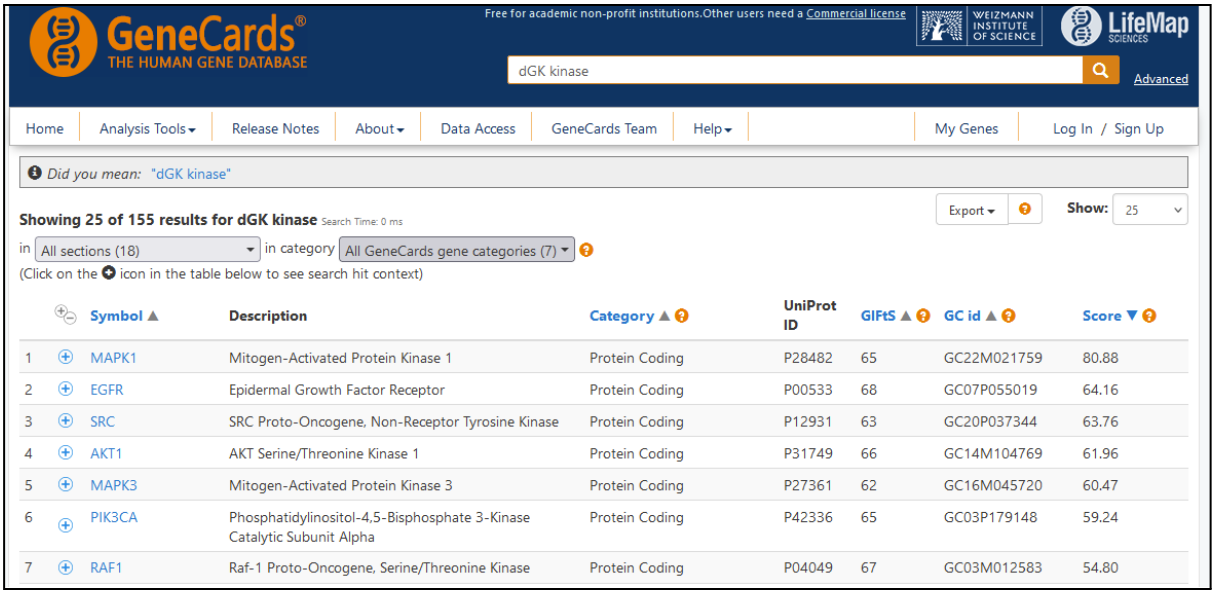
➔ **AmiGO2 - Gene Ontology**: AmiGO2 provides access to Gene Ontology (GO) terms, enabling detailed searches related to gene functions, processes, and cellular components. It is highly recommended for annotating gene-related entities.



➔ **PANTHER**: The PANTHER database provides gene ontology classifications, biological pathways, and protein-coding gene annotations. It can help annotate the functional aspects of gene-related entities.

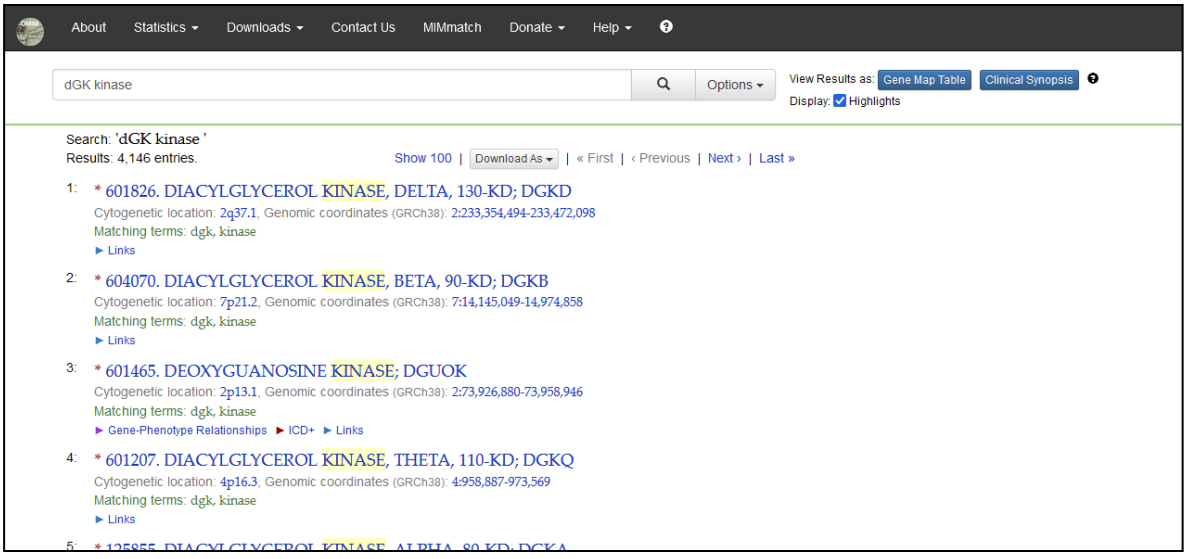


➔ **GeneCards**: GeneCards provides comprehensive information on human genes, including functions, pathways, and interactions. It is a useful tool for annotating genes and understanding their relationships.

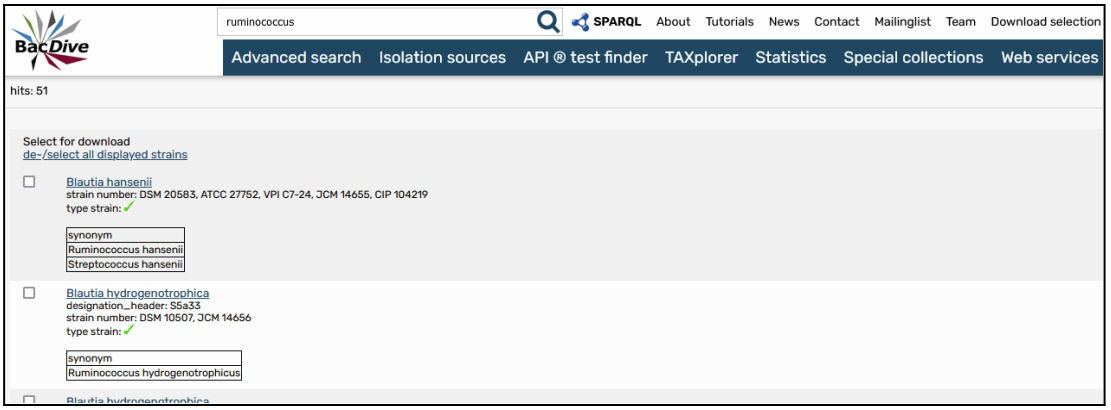


	Symbol	Description	Category	UniProt ID	GIFtS	GC id	Score
1	MAPK1	Mitogen-Activated Protein Kinase 1	Protein Coding	P28482	65	GC22M021759	80.88
2	EGFR	Epidermal Growth Factor Receptor	Protein Coding	P00533	68	GC07P055019	64.16
3	SRC	SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase	Protein Coding	P12931	63	GC20P037344	63.76
4	AKT1	AKT Serine/Threonine Kinase 1	Protein Coding	P31749	66	GC14M104769	61.96
5	MAPK3	Mitogen-Activated Protein Kinase 3	Protein Coding	P27361	62	GC16M045720	60.47
6	PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha	Protein Coding	P42336	65	GC03P179148	59.24
7	RAF1	Raf-1 Proto-Oncogene, Serine/Threonine Kinase	Protein Coding	P04049	67	GC03M012583	54.80

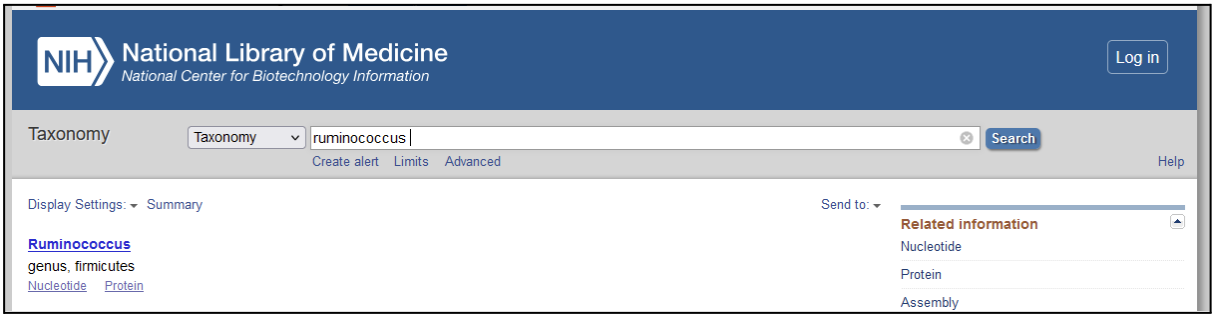
➔ **OMIM (Online Mendelian Inheritance in Man)**: OMIM is a comprehensive resource for human genes and genetic disorders. It is useful for annotating entities related to genetic and disease conditions.



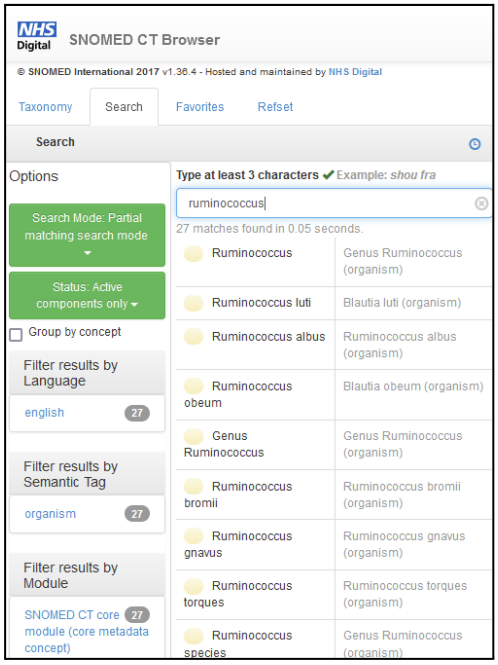
➔ **BacDive**: a metadatabase providing detailed information about bacterial diversity, including taxonomic classification. It is useful for verifying bacterial entities.



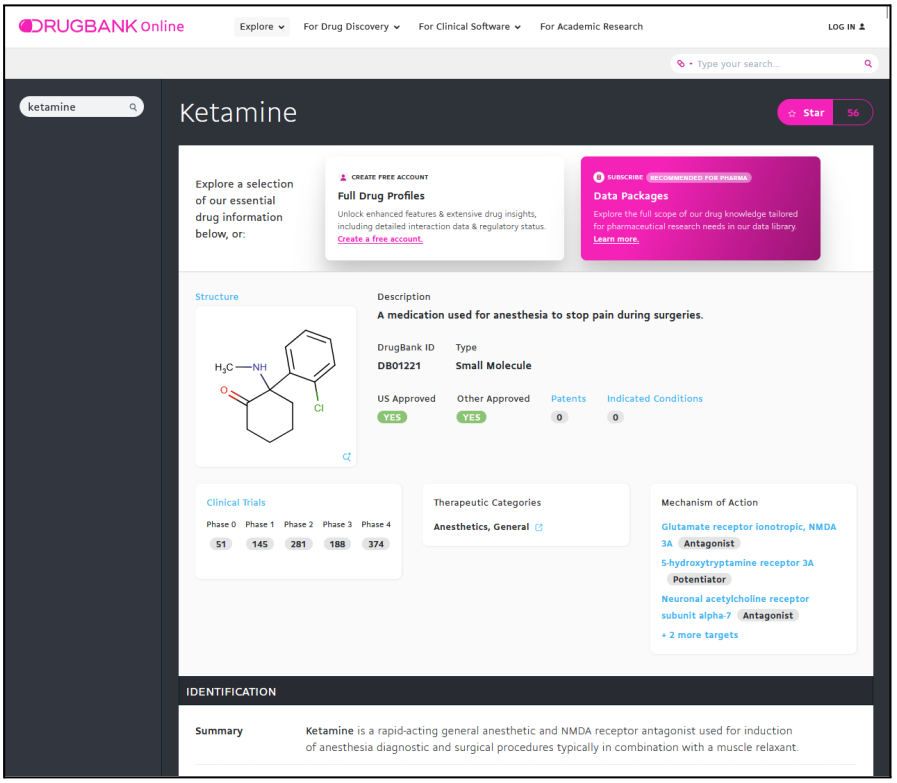
➔ [NCBI Taxonomy]: A curated classification of species, allowing annotators to verify biological names and classifications of bacteria, fungi, and viruses.



➔ [SNOMED CT Browser]: SNOMED CT is one of the most comprehensive clinical terminologies, covering a wide range of medical concepts. It is useful for verifying clinical and disease-related entities.



➔ **DrugBank**: DrugBank provides detailed information about drugs, including mechanisms, interactions, and targets. It is highly suitable for annotating drug-related entities.



➔ QuickGO: Gene Ontology and GO Annotations (TODO)

➔ Mouse Genome Informatics (TODO)

➔ NCBI Gene (TODO)

7. Bibliography

[1] Po-Ting Lai et al., “EnzChemRED, a Rich Enzyme Chemistry Relation Extraction Dataset,” *Scientific Data* 11, no. 1 (September 9, 2024): 982, https://doi.org/10.1038/s41597-024-03835-7.

[2] Ling Luo et al., “BioRED: A Rich Biomedical Relation Extraction Dataset,” 2022, https://doi.org/10.48550/ARXIV.2204.04263, see: BioRED_Annotation_Guidelines.

[3] Anastasia Krithara et al., “BioASQ-QA: A Manually Curated Corpus for Biomedical Question Answering,” *Scientific Data* 10, no. 1 (March 27, 2023): 170, https://doi.org/10.1038/s41597-023-02068-4.

[4] Ornella Irrera, Stefano Marchesin, and Gianmaria Silvello, “MetaTron: Advancing Biomedical Annotation Empowering Relation Annotation and Collaboration,” *BMC Bioinformatics* 25, no. 1 (March 14, 2024): 112, https://doi.org/10.1186/s12859-024-05730-9.

[5] C. -H. Wei, R. Leaman and Z. Lu, "SimConcept: A Hybrid Approach for Simplifying Composite Named Entities in Biomedical Text," in *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1385-1391, July 2015, doi: 10.1109/JBHI.2015.2422651

[6] Victor Sanh et al., “Learning from Others’ Mistakes: Avoiding Dataset Biases without Modeling Them” (arXiv, 2020), https://doi.org/10.48550/ARXIV.2012.01300.

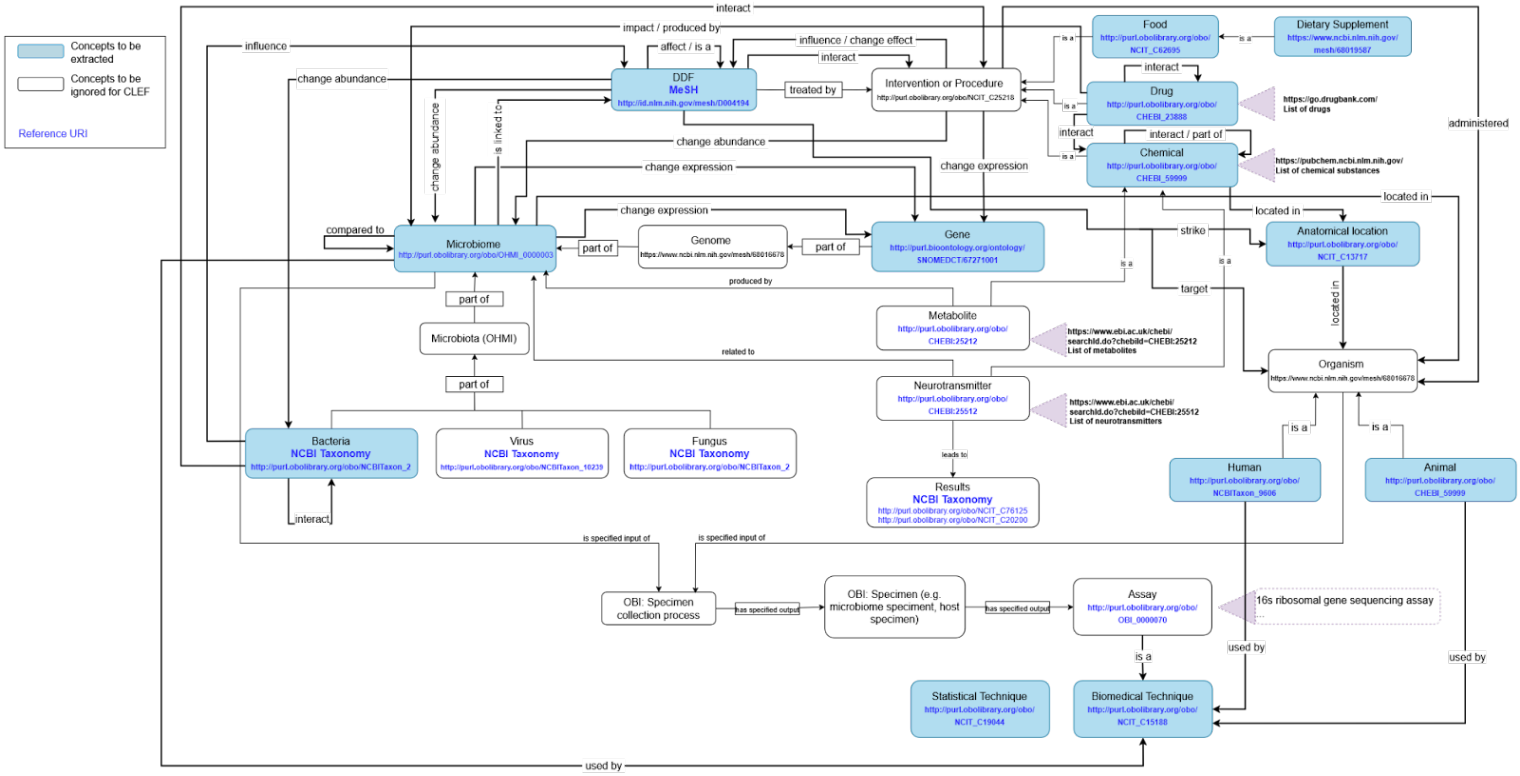
[7] John F. Cryan et al., “The Microbiota-Gut-Brain Axis,” *Physiological Reviews* 99, no. 4 (October 1, 2019): 1877–2013, https://doi.org/10.1152/physrev.00018.2018.

[8] Ting Liu et al., “Exploring the Microbiota-Gut-Brain Axis for Mental Disorders with Knowledge Graphs,” *Journal of Artificial Intelligence for Medical Sciences* 1, no. 3–4 (2021): 30–42, https://doi.org/10.2991/jaims.d.201208.001.

[9] Ting Liu et al., “Influence of Gut Microbiota on Mental Health via Neurotransmitters: A Review,” *Journal of Artificial Intelligence for Medical Sciences* 1, no. 1–2 (2020): 1–14, https://doi.org/10.2991/jaims.d.200420.001.

8. Appendix

Conceptual System



The draw.io version of this conceptual system is available at: [Draw.io Conceptual System](#).