GutBrainIE Annotation Guidelines

Anonymous authors

Document Structure

1. Overview

The GutBrainIE 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, while **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 mental health [7], extracted from PubMed. The entities and relations of interest are defined by the conceptual system that can be found 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 genes, bacteria, intermediates, and disorders.
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 using annotations created by a team of expert annotators and validated by a team of biomedical specialists.

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 two 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 types 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 label $y_i \in (y_1, y_2, \cdots, y_m)$. 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 type that can be found in the texts.

2. **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].

Normally, the process of dataset crafting also comprises the *Named Entity Normalization (NEN)* task, namely linking identified entity mentions to stable and unique database identifiers (e.g., MeSH, ChEBI) to standardize and contextualize entities [1]. At the moment we are not interested in this task.

Annotation process in practice

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

Before starting to annotate, annotators are required to **read the guidelines carefully**, paying particular attention to the defined entity and relation labels to ensure 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 Annotation**: Once familiar with the abstract, enable the visualization of the pre-annotated labels and proceed with NER. When an entity instance that appears multiple times in the document is annotated for the first time, the '**annotate all**' *MetaTron* feature may be used to uniformly annotate all subsequent occurrences of the entity, ensuring consistency throughout the document.
3. **RE Annotation**: After completing NER, proceed with RE to identify and classify relationships between the annotated entity mentions.

3. Entity Annotation Guidelines

3.1. Entity Labels

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

Entity Label	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 GutBrainIE 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 Label	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

Entity Label	URI	Definition	Reference URL	Notes	Examples
					acetate-producing bacteria, 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. --- This category also includes metabolites, which in biochemistry are the intermediate or end product of metabolism (more information here), and neurotransmitters, which are endogenous compounds used to transmit information across the synapses.	link	The list of chemicals reported in https://pubchem.ncbi.nlm.nih.gov/ can be used as a reference for this label.	The administration of sodium butyrate and cryptotanshinone (CPT) led to… We observed differentially abundant microbial-derived neuroactive metabolites including multiple B-vitamins, kynurenic acid, gamma-aminobutyric acid and short-chain fatty acids. Microbial metabolites (short-chain fatty acids -SCFAs-, bile acids, amino acids, tryptophan -trp- derivatives, and more), work as signaling pathways. The gut microbiota produces and modulates neurotransmitters such as GABA, serotonin, dopamine, glutamate, etc. Both Nef and Flu treatments induced significant increases in the levels of anti-depressant neurotransmitters, including dopamine (DA), serotonin (5-HT), and norepinephrine (NE).

Entity Label	URI	Definition	Reference URL	Notes	Examples
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. 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.
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).

Entity Label	URI	Definition	Reference URL	Notes	Examples
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.
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	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

Entity Label	URI	Definition	Reference URL	Notes	Examples
				of the abstract to determine whether the reference is to the gene or to the protein encoded by that gene.	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 discoveries. For instance, mentions such as "*psychiatrist*", "*clinicians*", "*medical personnel*", and similar terms should also be annotated under this entity label.	link	Instances of "*... population*", such as "*pediatric population*", should be annotated as human entity mentions. On the other hand, instances of "*... model*", such as "*MDD model*", should **NOT** be annotated.	Bipolar disorder is rare among populations that have not adopted contemporary Western lifestyles, which supports the hypothesis that bipolar disorder results from a 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*,	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

Entity Label	URI	Definition	Reference URL	Notes	Examples
				at the current stage, we want to annotate both *microbiomes* and *microbiota* as *microbiome*.	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.

3.2. General Entity Annotation Rules

➔ **Annotate all concept types**

- Annotate all text spans corresponding to the thirteen defined concept types [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 *<i>*, *<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 have only one entity label. If an entity could belong to multiple categories, use the context within the sentence to determine the appropriate label.

- 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 *neurotransmitter*, therefore as a *chemical*.
- In some cases, it can be challenging to determine whether an entity mention should be annotated as a *metabolite*, *gene*, *neurotransmitter*, or *chemical*. This difficulty arises because these categories are all subsets of *chemical*, meaning that metabolites, genes, and neurotransmitters are also chemicals. When the appropriate label cannot be determined through annotation tools, context, or reasoning, the entity should be labeled with the most generic category *chemical*. Examples:
    - In “the enzyme was inhibited by 3-hydroxybutyrate” it is not clear by the context if the entity mention is being used as a *metabolite* or a *chemical*. Therefore, it should be annotated as a *chemical*.
    - In “ASD patients shower increased dopamine compared to…” it is not clear by the context if dopamine is being used as a *neurotransmitter* or as a *chemical*. Therefore, it should be annotated 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."

- 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 "*xxx transplant*", "*xxx therapy*", "*xxx treatment*", "*xxx intervention*", and "*xxx scale*" are general medical procedures,

therapeutic approaches, or assessment frameworks and therefore should not be annotated as *biomedical technique* entity mentions.

- 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.

➔ **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]. 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. Relation Annotation Guidelines

4.1. 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**

4.2. 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.
- 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” (*DDF*), and the other from “probiotics supplementation” (*dietary supplement*) to “depression” (*DDF*).
- 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” (*DDF*) 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” (*DDF*), 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)” (*DDF*) 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)” (*DDF*).
- 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” (*DDF*) through its production of “gamma-aminobutyric acid (GABA)” (*chemical*). Therefore, an *influence* relation can be inferred between "Bacteroides fragilis" (*bacteria*) and "anxiety" (*DDF*).

➔ **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" (*DDF*), and a second one from "SSRIs" (*drug*) to "depression" (*DDF*).

➔ **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.

5. 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.

6. Appendix

Conceptual System