Annotation Guidelines for *Chemical Entity Mentions for Assessment of MTI* (*ChEMFAM*) Corpus

Contents

[Introduction 2](#_Toc16070593)

[Organics 2](#_Toc16070594)

[Inorganics 2](#_Toc16070595)

[Amino acids, peptides, and proteins 3](#_Toc16070596)

[Nucleic acids, nucleotides, nucleosides 3](#_Toc16070597)

[Nesting 4](#_Toc16070598)

[Fragments 4](#_Toc16070599)

[Stop words 4](#_Toc16070600)

[Miscellaneous 5](#_Toc16070601)

# Introduction

This document records the guidelines used to annotate chemical entities in a set of 200 MEDLINE citations, consisting of abstracts and titles, referred to as the Chemical Entity Mentions for Assessment of MTI (ChEMFAM) corpus. We have chosen to divide chemicals that appear in this corpus into four categories, corresponding to MeSH tree terms:

1. Inorganic Chemicals (D01)
2. Organic Chemicals (D02)
3. Amino Acids, Peptides, and Proteins (D12)
4. Nucleic Acids, Nucleotides, and Nucleosides (D13

Entities that represent an instance of a chemical are annotated. For the purposes of this collection, an instance of a chemical is defined as an entity that can be mapped to IUPAC nomenclature, a gene, or a protein expressed by a single gene or formed of multiple units expressed by multiple genes. Entities that represent chemical classes or general concepts that encompass many chemicals are not annotated. These entities are added to a stop word list.

The rest of this document describes in further detail the guidelines for annotating entities belonging to each category, adding entities to the stop word list, making nested and fragmented annotations, as well as a miscellaneous section covering general guidelines not mentioned in the other sections. For brevity, the Amino Acids, Peptides, and Proteins class is referred to as proteins, and the Nucleic Acids, Nucleotides, and Nucleosides class if referred to as genes.

# Organics

1. Compounds with C-H bonds are considered organic and are annotated.
2. Chemicals historically considered organic, such as CCl4, are annotated.
3. All IUPAC nomenclature is annotated.
4. Any common chemical names that can be mapped to an IUPAC chemical are annotated. See 8 for the exception.
5. Chemical formulas are annotated.
6. Ions not found within larger organic entities are annotated.
7. Any structural or functional classes of organic chemicals are added to the stop word list.
8. Brand name drugs are not annotated nor added to the stop word list.
9. Generically named drugs are not annotated unless the article specifically discusses the chemical properties, mechanisms, or chemical interactions of the drug.
10. If an isomer of a chemical is referred to without the chemical name, it is added to the stop word list, e.g., (-)-R-enantiomer
11. Functional groups are added to the stop word list.

# Inorganics

1. Compounds without C-H bonds are considered inorganic and are annotated.
2. All IUPAC nomenclature is annotated.
3. Any common chemical names that can be mapped to an IUPAC chemical are annotated.
4. Inorganic compound formulas are annotated.
5. Ions not found within larger inorganic entities are annotated.
6. Very common chemicals are not annotated unless the article specifically discusses the chemical properties, mechanisms, or interactions of the chemical. In this collection, this only applies to water.
7. Nanoparticles or other materials are not annotated.
8. Formulas of inorganic materials are not annotated, only the individual elements.
   1. thermoluminescence sensitivity in LiF:Mg,Cu,Ag material
   2. The temperature-insensitive NaYF4: Nd nanophosphors
9. Minerals are annotated.
   1. Brushite
   2. Zirconium

# Amino acids, peptides, and proteins

1. Any specific factors, receptors, hormones, other types of proteins are annotated. As a general rule, if there is a single gene that codes for the entity or if the entity is composed of multiple components coded for by multiple genes, that entity is annotated.
2. Amino acids, peptides and other molecular structures composed of amino acids are annotated.
3. In general, if a named gene, transcript, or other gene product can be mapped to an entry in an online database, i.e., it has been researched and named, it is annotated.
4. Functional, structural, or other types of classes are not annotated. These are be added to the stop word list.
   1. Receptor
   2. inhibitor
   3. transporter
5. Classes of proteins are not annotated. However, occasionally the authors of an article will refer to an instance of a protein by using its more general class name. This is particularly common when the researchers are measuring the activity of the protein. In these cases, the protein entity is annotated. However, if the article mentions the class and an instance, only the instance is annotated.
6. If the class of the chemical (protein vs gene) cannot be determined from the context of the article, the entity is annotated as a protein.
7. If the method used for expression detection is western blot or another protocol used to quantify protein, assume all references to entities are protein, unless specifically mentioned otherwise.
8. No mutations or locations are annotated or added to the stop word list, explained further in the nucleic acids section, e.g., Trp19-Arg124.

# Nucleic acids, nucleotides, nucleosides

1. Any instance of a gene, nucleotide, nucleic acid, nucleoside, DNA or RNA, mRNA, miRNA, shRNA, etc, is annotated. While ribonucleoside-based entities are chemically distinct from nucleoside-based entities, we do not require this distinction for the purposes of MEDLINE indexing.
2. In general, if a named gene, transcript, or other gene product can be mapped to an entry in an online database, i.e., it has been researched and named, it is annotated.
3. In general, if the entity is transcribed or part of something that is transcribed, it is labeled in this category.
4. Genotypes in format AA/BB are not annotated. However, since animal lines can be described by a single gene, these genes are annotated.
   1. Mdx mice.
   2. dysf(-/-)
5. Entities specifying a base or mutations at a specific location are not annotated. These are also not added to the stop word list.
   1. nonsense variant (Lys177X)
   2. polymorphism of MTHFR C677T/A1298C
   3. heterozygous genetic variant was detected in exon 4, position 1027 (C1027T)
6. Functional, structural, or other types classes are not annotated. These are added to the stop word list.
   1. Transcript
   2. Intron, exon
   3. DNA
7. Abbreviations of bases are not annotated. These are not added to the stop word list.
   1. A
   2. G
   3. AG
   4. TT

# Nesting

1. Organics and Inorganics
   1. Entities that only consist of organic or inorganic chemicals are not annotated with nested annotations. An exception was made for the entity *copper hydroxide nanozyme*, as this entity is syntactically and functionally structured like a protein.
   2. If an organic or inorganic chemical is found within another type of entity (genes or proteins) to be annotated, the longest possible organic or inorganic nested within the larger entity is annotated.
2. Proteins and Genes
   1. All entities within an entity, except stop words, are annotated
   2. Protein or genes nested within a parent entity are annotated with the class of the parent entity, unless very clear that the child is not the class of the parent. Organic or inorganic entities within a parent entity are assigned the appropriate class.
   3. Nested entities are only created if a dash or other symbol separates the entities within the parent entity.

# Fragments

1. All categories can be annotated with fragments.
2. When multiple chemicals with a common ending or beginning are listed in the text, the fragments feature in the brat annotator is used to annotate these terms, e.g., oleic, linoleic and linolenic acids
3. When an acronym is located within the middle of a entity to be annotated, a fragment is added around the acronym, and the acronym is annotated separately, e.g., Interleukin (IL)-6

# Stop words

1. All entities used to describe a chemical entity without including information regarding structural, mechanistic, or other chemical properties are added to the stop word list.
   1. agonist
   2. hormone
   3. adduct
   4. catalyst
2. Entities that describe functional, structural or other property-based classes as added to the stop word list.
3. Often multiple variants of a single gene, gene product, or protein exist. These are sometimes denoted with alpha, beta, or roman numerals. If the authors fail to specify the exact instance, and are clearly not referring to a specific instance in the context of the text, the entity is added to the stop word list.

# Miscellaneous

1. The parentheses around acronyms are not annotated.
2. If a chemical is part of a compound adjective or noun, the chemical within the larger compound entity is annotated, e.g., propylene oxide-induced.
3. If a chemical appears in a heterogeneous substance, microscopic substance, or material, only the chemical/s is annotated and not the entire entity.
4. If a chemical is part of non-chemical acronym, the letters in the acronym that correspond to the chemical are not annotated, if there are no dashes or other symbols separating the letter representing the chemical from the rest of the letters in the acronym, e.g., fasting plasma glucose (FPG).
5. If an entity is annotated, the nested stop words are not added to the stop word list. They may be on the stop word list, if that stop word appears elsewhere, outside of an annotated entity.