CHEMDNER-patents: Chemical Entity annotation manual

Version 2.0 (May 2015)

This document describes the annotation guidelines used for the construction of the annotations chemical mentions (CEM – Chemical Entity Mention) of the CHEMDNER-patents corpus. These annotation guidelines were generated based on the CHEMDNER 2013 task exclusively focused on chemical entities extracted from papers. Thus, annotation guidelines were adapted to mine patents with a stronger focus on identifying any wide definition of chemical terms rather than in obtaining highly refined chemical mentions that can be univoquely translated into a chemical structure. The reason for this subtle refinement resides in the fact that patents contain very wide, general chemical mentions to describe the different substituents of the general Markush formula. Discarding this kind of mentions would prevent future applications (e.g. deconvolution and interpretation of Markush formula) of the CHEMDNER-patents corpus. However, no attempt is made to establish Markush structure relation: establishing meaningful correlations between the Markush formula and substituents. Thus, two related CEMs that are clearly separated as different entities in the text will be annotated as two different CEMs; although the meaning of one of them can be linked to the other one.

This guide provides the basic details of the CEM task and the conventions that should be followed during the corpus construction process. The annotation guidelines were refined after iterative cycles of annotations of sample documents based on direct suggestions made by the curators as well as through the observation of inconsistencies detected when comparing the results provided by different curators. Some participating teams provided feedback to improve the documentation after the release of the first sample set prepared for the CHEMDNER 2013 task. These informal rounds of curation served to improve the guidelines in the sense of making them more intuitive and easy to follow for the annotators.

The manual annotation task basically consists in labeling or marking up manually through a customized web-interface (AnnotateIt) the mentions of chemical entities. This was done following a set of rules that will be specified in more detail below. The text to be labeled consisted of patent abstracts (titles and abstracts in English) from patents published between 2005 and 2014 that had been assigned to the IPC codes A61P and A61K31.

The selected chemical entity mentions were classified into one of seven chemical entity mention (CEM) classes defined in more detail below. The color code corresponds to the color tags provided by the annotation interface for each of the CEM classes, to make the manual labeling and visualization easier.

CEM class	Description	Examples
SYSTEMATIC	Systematic names of chemical mentions, e.g. IUPAC and IUPAC-like names.	2-Acetoxybenzoic acid 2-Acetoxybenzenecarboxylic acid 2-Acetoxybenzoic acid N-(4-hydroxyphenyl)acetamide 3,5,4'-trihydroxy-trans-stilbene
IDENTIFIERS	Database identifiers of chemicals: CAS numbers, PubChem identifiers, registry numbers and ChEBI and CHEMBL ids	CAS Registry Number: 501-36-0445154 CID 445154 CHEBI:28262 CHEMBL504
FORMULA	Mentions of molecular formula, SMILES, InChI, InChIKey	CC(=O)Oc1ccccc1C(=O)O InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12) C9H8O4 (CH3)2SO
TRIVIAL	Trivial, trade (brand), common or generic names of compounds. It includes International Nonproprietary Name (INN) as well as British Approved Name (BAN) and United States Adopted Name (USAN)	Aspirin Acylpyrin paracetamol acetaminophen Tylenol Panadol resveratrol
ABBREVIATION	Mentions of abbreviations and acronyms of chemicals compounds and drugs	DMSO GABA
FAMILY	Chemical families that can be associated to some chemical structure are also included. It involves: i-FAMILY- SYSTEMATIC: IUPAC (plurals) ii-FAMILY- FORMULA iii-FAMILY- TRIVIAL iv.-FAMILY- ABBREVIATION v- FAMILY – FAMILY	Iodopyridazines (FAMILY- SYSTEMATIC) diphenols (FAMILY- SYSTEMATIC) quinolines (FAMILY- SYSTEMATIC) terpenoids (FAMILY- TRIVIAL) ROH (FAMILY- FORMULA)
MULTIPLE	Mentions that do correspond to chemicals that are not described in a continuous string of characters. This is often the case of mentions of multiple chemicals joined by coordinated clauses.	thieno2,3-d *and* thieno3,2-d *fused* oxazin-4-ones

Table 1. Chemical Entity Mention (CEM) classes defined for the CHEMDNER-patent task. For each CEM a short description and illustrative example cases are provided.

2. CHEMDNER chemical entities

The focus for defining the chemical entities annotated for the CHEMDNER-patent task was primarily to capture those types of mentions that are of practical relevance from patents' perspective. Therefore the covered chemical entities had to represent those kinds of mentions that can be exploited for extracting chemical structural information from patents.

The annotation carried out for the CHEMDNER-patent task was only exhaustive for the types of chemical mentions that are described in more detail below. This implies that other types of mentions of chemicals and substances were not labeled. The common characteristic among all the chemical mention types used for the CHEMDNER-patent task was that i) they could be associated to chemical structure information to at least a certain degree of reliability or ii) they could be associated to general chemical structural information according to the terms commonly found in patents, specially those describing the characteristics of the substituents of the Markush formula (e.g., "heteroaryl and aromatic bicycles" that describe topological classes). This implied that, compared to the CHEMDNER 2013 task (focused on scientific articles), an extended range of general chemical concepts (e.g., describing topologic features – see rule N3) were included. However, general chemical concepts (non-structural or non-specific chemical nouns), adjectives, verbs and other terms (reactions, enzymes) that cannot be associated directly to a chemical structure are excluded from the annotation.

The annotation process itself also relied heavily on the domain background knowledge of the annotators when labeling the chemical entity mentions. A requirement to carry out the manual annotation was that annotators should have a background in chemistry, chemoinformatics or biochemistry to make sure the annotations are correct. This also made it possible to provide a short and compact set of annotation rules rather then requiring very detailed guidelines for non-experts. In this sense we followed a similar strategy as done for the gene mention tasks of previous BioCreative efforts (Smith et al. 2008). The definition of the chemical entity mention types used for the CHEMDNER task were inspired by the annotation rules used by Kolaric et al. (2008) and by Corbett et al. (2007).

Chemical Entity Mentions (CEMs) for this task had to refer to names of specific chemicals, specific classes of chemicals or fragments of specific chemicals. General chemical concepts, proteins, lipids and macromolecular biochemicals are excluded from the annotation. Therefore genes, proteins and protein-like molecules (> 15 amino acids) were excluded from the annotation. Chemical concepts were annotated only if they provided structural information (e.g. FAMILY type detailed below).

In order to label chemical entity mentions a set of rules have been defined that are described below. Example cases are provided to aid in understanding the different rules. The correct CEM cases are marked in gray.

As first general annotation guidelines consider:

Rule 1 → Use of external knowledge sources

In case the curator is not sure if a mention corresponds to a compound or he does not know what kind of compound mention it is, he may consult external knowledge resources: Wikipedia, Chemspider, Chemical Suppliers Catalogues (Sigma Aldrich, Tocris,…), Scifinder , http://global.britannica.com/ such as the web or chemical databases to resolve doubts. A list of useful external knowledge sources should be compiled. Ideally some aid here from the annotation system should be expected.

Rule 2→ Not unclear mentions

Do not tag unclear cases. If the annotator is not sure about a given mention, even after consulting some external sources, the corresponding mention should remain unlabelled.

The following examples are just to exemplify hypothetical doubtful situations:

Alkaloid *stands for compounds with a basic nitrogen, but the boundary is not clear enough and the substructural pattern neither. However, chemists typically recognize them, so this term should be labelled.*

Glucocorticoid *a specific type of steroid hormone classified according to the receptor to which they bind (glucocorticoid receptor). Although there is a biological role in the chemical name, the general structure of "glucocorticoid" can be devised as a steroid.*

The following annotation rules define which chemicals are CEM

Positive Rules – CEM are:

P1. Chemical Nouns convertible to:

-A single chemical structure diagram: single atoms, ions, isotopes, pure elements and molecules:

Fluorine, Iron, Deuterium, Benzene, Pyridine

-A general Markush diagram with R groups. Typically, chemical functionalities, fragments and structural classes → assignable to the CEM = FAMILY class.

Amides, Hydroxipyridines, ROH, Amino acids, Methyl Group, O-H group

P2. General class names where the definition of the class includes information on some structural information or elemental composition, independently of their origin (synthetic small compounds or natural products) → CEM = FAMILY class

Hydrocarbons, organochlorines, carbohydrates, organometallics, Lewis Acids, Grignard Reactants, polyketides, steroids, macrolides, terpenoids, fatty acids, nucleotides, nucleobases, Bronsted-Lowry acid, transition metal, halogen, Schiff base, Wittig Salt, Wittig Reagent, monosaccharide, sugars, saturated fatty acids, trans fatty acids, triglyceride, bile acid, bile salt, statin, betaine, ceramide, glucosylceramide, carotenoid, ketocarotenoid, alkaloid, androgen, estrogen, aglycone, cannabinoid, opioid, phosphoinositide, saponins, sapogenin, parabens, vitamin, mineralocorticoid, alkane, alkene, alkyl, leukotriene, steroid hormone, glucocorticoid, glycosphingolipid, sphingolipid, aminoglycoside, glycoside, oligosaccharide, dipeptide, nonapeptide, glucoside, nucleotide, glucosinolate,rare earth, amino acid…

Note: in the case of biochemical terms (mostly general mentions of natural products) that refer to general classes encompassing small molecules as well as larger entities that are in the frontier of macromolecules (see P3 and N5), the entity should preferably be labeled. For example, *oligosaccharide* term refers to molecules consisting of a small number of simple sugars; typically three to nine. As stated below under rule P3, saccharids with up to three sugars (trisaccharides) are annotated. Thus, the term oligosaccharide should be annotated as it contains, among others, the trisaccharides. A similar case for glycoside, ceramide…

P3. Small Biochemicals

-Saccharids: monosaccharides, disaccharides and trisaccharides should be tagged:

Glucose	*monosaccharide*
Fructose	*monosaccharide*
Ribose	*monosaccharide*
Sucrose	*disaccharide*
Streptomycin	*an aminoglycoside trisaccharide*
Gentamicin	*an aminoglycoside trisaccharide*
Cyclodextrin	*cyclic oligosaccharides*; *not tagged*

-Peptides and proteins: peptides and peptidomimetics should be tagged. By convention, a threshold of 15 amino acids was chosen as cut-off. Thus, peptides with less than 15 amino acids should be tagged as CEM (both, cyclic and non-cyclic peptides).

Glutathione	*trimer*
Cyclosporin A	*11 amino acids*
Degarelix	
Gonadotropin-releasing hormone (GnRH)	*with 10 amino acids*
Luteinizing-hormone-releasing hormone (LHRH)	*same case as for GnRH*
Gonadotrophin-inhibitory hormone (GnIH)	*with 12 amino acids*
Azaline B	*with < 15 amino acids*
Angiotensin	*<10 amino acids*
Gramicidin	*15 amino acids*

As well as chemical modifications on these peptides:

[D-Ser-(But),6, des-Gly-NH210]LHRH ethylamide

But, for example, luteinizing hormone is a protein (92 amino acids), so it should not be tagged. In the same way, chemically modified specific proteins with > 15 amino acids should not be tagged to avoid tagging alone a chemical modifier attributable to a non-tagged specific entity:

Luteinizing hormone (LH) *untagged because it has 92 amino acids*

Acetylated insulin

leave acetylated untagged as insulin is untagged

-Nucleotides: Mentions of monomers, dimers, trimers should be tagged.

NADH
NAD+
Nicotine adenine dinucleotide
ATP
Adenosine Triphosphate
Adenosine 5'-Triphosphate
SAM
S-Adenosyl methionine
cAMP

but not polysaccharides: Polysialic acid, starch, cellulose, PolySia

In the same way, chemically modified specific biopolymers should not be tagged to avoid tagging alone a chemical modifier attributable to a non-tagged specific entity:

Hydroxypropylmethyl cellulose

as cellulose is explicitly not considered as a CEM, do not annotate the chemical modifier preceding the name.

-Lipids: Fatty acids and their derivatives (including tri-, di-, monoglycerides), sterol derivatives…excluding polymeric structures

Glycerol
Prostaglandin A
Leukotriene A4
Cholesterol
Lipopolysaccharides
Eicosanoide

P4. Synthetic Polymers

Nylon
Polystyrene
Sulfonate polystyrene
Polyvinyl chloride (PVC)
Polyamides
Polyacrylamide (PAM)
Nafion

Note: synthetic polymers that are composed of biological monomers (synthetic biopolymers), should also be annotated:

PolyHis
PolyAsp
Poly(l-lactic acid)
Peptoid compounds

P5. Special Cases

-Minerals:

Calcite
Silica
Alumina
Titania

-Laboratory Reagents: common synthetic chemistry laboratory reagents, but only if their chemical composition is well defined

Petroleum ether
Silica gel
Universal indicator
Molecular Sieves
Litmus
Paraffin oil
Saline

-Dye and indicator names:

methyl red
Coomassie Brilliant blue
DAPI

Negative Rules – CEM are not:

N1. **Other terms different from chemical nouns**: adjectives (if isolated/outside from chemical nouns – see M3 and M4 below), pronouns, verbs, other terms (reactions and enzymes), chemical prefixes (if isolated/outside from chemical nouns), anaphors, referring expressions, compound numbers…

- Chemical Reactions:

Deshidrogenation
methylation
hydrolysis

- Pronouns, anaphors:

"DAPI is a dye…**this** compound…" *do not tag "this"*

- Compound numbers in anaphors: Even if the numbers are combined with other word (generating anaphors), they should never be annotated:

…of 8-amino-2,6-methano-3-benzazocine (2)…	*do not tag "2"*
(S)-4-AHCDP (6) and (R)-4-AHCP (7)	*do not tag "6" and "7"*
cis-9, ortho-12	*do not tag these entities*

- Chemical Prefixes (outside chemical names):

1,4-derivatives *do not tag "1,4-"*

N2. Chemical nouns named for a role or similar, that is, nonstructural concepts:

- **Generalities**: analogue, substituent, inhibitor, hit, agonist, antagonist, activator, effector, antioxidant, substrate, inactivator, pigment, agent, standard, pharmacophore, drug, promoter, exon, intron, gen, antifolate, food, compound,…

But for example annotate particular names that contain any of these words. For example:

Compound C *corresponds to a chemical substance (cas 866405-64-3)*

- **Biological Roles**: hormone, antibiotics, antigen, herbicides, antifungals, toxin, metabolite, antineoplastic agents, antiestrogens, …

But for example annotate particular cases that can be connected with a general or particular chemical structure. For example:

Ciguatoxin	*family of toxins*
Aflatoxins	*family of toxins composed of difuran* and *coumarin*
Aflatoxin B1	*particular chemical compound*
tyroid hormone	*corresponds to two well-defined chemical compounds*

- **Industrial Roles**: Surfactant, Plasticizer, Aerolizer

- **Reactivity Role**: electrophile, nucleophile, michael acceptor, dienophile, chelator, alkylating reagent, oxidizer, cation, anion, lipophile, zwitterion …

But annotate:

- particular cases in which some of these words (cation, anion, zwitterion) are part of a longer specific chemical name (see M2).

fluoranthene anion	*a particular anion*

- particular cases in which the adjective form of these words adds additional information on the chemical structure (see M3).

anionic PAM	*specific subclass of polyacrylamides*
anionic MF1011	*a particular polyacrylamide*
zwitterionic siloxane	*adds chemical information*

- **Laboratory Role**: solvent, reagent, starting materials, building blocks, buffer, catalyst, solution

ammonia solution	*only ammonia is annotated*

But annotate CEM's in which the word "block" has chemical sense (see M1)

polyethylene glycol-block-poly(d,l-lactide)

- **Elementary Particles**: neutron, proton, electron, helion

- **Plants** (and APIs from plants without a defined chemical structure): estragon, bromelain, silymarin

- **Oils, essences and general formulations of several compounds**: estragon

- **Antibodies or related**: trastuzumab, alemtuzumab, Brentuximab vedotin, Daclizumab, denosumab, herceptin, Kadcyla, MEDI-575, Natalizumab, Ocrelizumab

N3. Very nonspecific structural concepts:

- **General structural concepts**: atom, ion, molecule, polymer, stereoisomer, enantiomer, isomer, conformer, mesomer, conformation, monomer, dimer, trimer, tetramer, lipid, gen, protein, alkali, functional groups, carrier proteins, saturated fat, terpolymer, …

The stereoisomer 6, but not 7, activated cloned	*not tagged*
Terpolymer	*not tagged*

But annotate:

- particular cases in which some of these words (ion, dimer, trimer) are part of a longer specific chemical name (see M2).

  chloride ion — *chloride as an ion (not as a gas)*
  thiol dimers — *dimers adds information on the thiol type*

- particular cases in which the adjective form of these words adds additional information on the chemical structure (see M3).

  Polymeric triamcinolone acetonide

- particular names that contain any of these words

  Lipid A — *corresponds to a chemical substance*
  Alkali metals

- **Vague topological descriptors**: macrocycle, catenane, rotaxane,…

But annotate:

- particular cases in which some of these words (macrocycle) are part of a longer specific chemical name (see M2)

  tetrapyrrole macrocycle

- topological terms that are part of popular fragment descriptions in chemical patents when listing the covered substitution patterns of the Markush formula. These topological terms will be assigned to the FAMILY class (or MULTIPLE in some cases).

  aromatic and heteroaromatic bicyclic compounds
  spiro-tetracyclic ring compounds
  8- to 11-membered bicyclic
  5-10-membered monocyclic or bicyclic aromatic ring — MULTIPLE

N4. Context Criteria: Words are not CEM if they are not CEM in context, even if they are co-incidentally the same set of characters (synonyms and metaphors):

- Lead compounds are often found in high-throughput screenings ("hits") or are secondary metabolites from natural sources → *not tagged*

- Mutations in ICE genes disrupting mating-body formation lead to 5-fold decreased ICE transfer rates → *not tagged*

- Lead is a chemical element in the carbon group with symbol Pb.

- The man without self-reliance and an iron will is the plaything of chance → *not tagged*

- What the new gold standard will look like → *not tagged*

- Phosphor → *not to be tagged when it refers to the luminescent substance*

N5. Biomolecules/Macromolecular biochemicals: not large oligomeric and polymeric or established DNA/RNA/protein sequences:

Do not tag proteins, polypeptides (> 15aa), nucleic acid polymers, polysaccharides, oligosaccharides (with > 3 monomers) and other biochemicals. Exclude all large biopolymers regardless of how their structures are organized. *Chemical*: if it is best represented using a chemical structure. *Biochemical*: if it is more usually represented using a sequence or a block diagram.

ubiquitin, insulin, DNA, mRNA, keratin, collagen, starch, cellulose, glycogen, agarose, chitin, murein, peptidoglycans, glycoproteins, lipopolysaccharide, interferon, human fibroblast interferon, Kozak sequence (*example of an established sequence of amino acids*), annexin, atrial natriuretic peptide (*28 amino acids*), peptide, ribonucleic acid, chitosan, nucleic acid, polynucleotide,

N6. General vague compositions

Pigments with a relatively varying mixture: melanin, vaseline, lanoline

N7. Special words not to be labeled by convention

Organic
Inorganic
Water and its physical states (Steam, Ice…) as well as adjectives (aqueous)
Proton, helion (proton for either the fundamental particle or the H+)

Note: Compared with the CHEMDNER 2013 task and following participant's suggestions, we decided to include special words LEAD and GOLD to be annotated (depending on the context) instead of automatically discarding them (as in CHEMDNER 2013). Thus, if the words *lead*, *gold*, *iron* refer to the chemical element (see N4), they should be tagged.

lead(II)
gold xantoghenates
lead(II) oxybromide
gold-thioglucose

3. CHEMDNER entity mentions type description

The following CEM types were annotated for the CHEMDNER corpus. The following general guidelines should be applied when annotating the different CEM types:

Rule 3→ Each chemical mention can only be marked as a single CEM type

Rule 4→ Priority rules of CEM of various types

In case a CEM is comprised of a combination of different types or mentions, e.g. systematic, trivial, abbreviation, etc, the curator should label the mention according to the ranking provided for the CEM, CEM1,… CEM7. For example, if it contains at least a part that follows IUPAC rules, it should be tagged as SYSTEMATIC (even if the rest of the mentions correspond to trivial names, formula or identifiers and the IUPAC string is relatively short).

Asp-Glu-NSP *FORMULA: where NSP is an abbreviation in the text*

Testosterone *TRIVIAL*
3H-Testosterone *SYSTEMATIC (as 3H is IUPAC)*

Sildenafil *TRIVIAL*
N-methyl sildenafil *SYSTEMATIC (as N-methyl is IUPAC)*

[N(gamma)-(IGly)Dab(8)]degarelix

N(gamma) is IUPAC so it is composed of IUPAC + formula + trivial → results in SYSTEMATIC

[(2-pyridyl)-methyl)d-Dap(3)]degarelix
IUPAC + Formula + Trivial → results in SYSTEMATIC

[IOrn(8)]degarelix
composed of Formula + Trivial → results in FORMULA

[Pra(7)]degarelix
composed of Formula + Trivial → results in FORMULA

CEM-1 (SYSTEMATIC): includes multi word systematic, CAS-style names and semi-systematic names such as mentions of chemical compounds following the IUPAC nomenclature guidelines (http://www.iupac.org/fileadmin/user_upload/publications/recommendations/CompleteDraft.pdf). Also IUPAC-like mentions are included, as often the authors do not follow strictly the guidelines and sometimes authors combine chemical mentions that have both systematic and non-systematic mention elements.

1,2-dimethyl-3-hydroxypyridin-4-one
acetone semicarbazone

1-octanol
chloroacetyl chloride
iron
sodium
iron(III)
iron(3+)
acetylsalicylic acid
ethyl group
methyl substituent

NOTE: mentions of unique substances (not general family compounds) that are IUPAC or IUPAC-like next to non-essential parts of the chemical entity or name modifiers (see M1, M4 and M9) should be assigned to the FAMILY class. This is specially important in patents as families of compounds are commonly written as the combination of a single "head" compound (pyridine,…) followed by the terms derivatives, analogues…

2,3-Dihydrobenzofuran analogues	*FAMILY (FAMILY-SYSTEMATIC)*
5-(1-azidovinyl)uracyl derivatives	*FAMILY (FAMILY- SYSTEMATIC)*

CEM-2 (IDENTIFIERS): corresponds to the following database identifiers of chemicals (strictly these databases): CAS registry numbers, PubChem, ChEBI and CHEMBL database identifiers and also company codes. These identifiers should only be labeled if the context provides enough information that these mentions correspond to chemical identifiers.

Its CAS Number is 28718-90-3…
Company codes: PD-0332991, FE200486

Note: if the CAS Number is written as "CAS # 2634-33-5" → only the numerical code should be written.

CEM-3 (FORMULA)*:* this class corresponds to mentions of chemical formula, chemical line annotations, SMILES, InChI and 3-letter codes of nucleotides, amino acids and monossacharides:

$C_6H_{12}O_6$	
EtOAc	
Fe,Na, Fe(III),Li^+, Fe^{2+}	*Atomic elements*
CC(=O)C	*Chemical Line annotations*
Glu-Cys-Gly	*3-letter codes of small peptides*
GlcNAc	*Oligosaccharides nomenclature: formula with abbreviation*
Asp-Glu-Fmoc	*Formula (formula with abbreviation)*
t-BuOK	

InChI=1S/C22H15N/c1-3-8-16(9-4-1)21-19-13-7-12-18-14-15-20(23(18)19)22(21)17-10-5-2-6-11-17/h1-15H

Note: valid 1-letter codes of amino acids (see O6) are also to be labeled as FORMULA, considering that the final sequence has a length < 15 amino acids.

Arg-Lys-Phe (RKF)	*both CEMS annotated as of type FORMULA*

CEM-4 (TRIVIAL): this class included trivial and common names of compounds. It also includes trademark and commercial names of chemicals and drugs.

-**Drug Names**: aspirine, Viagra, Degarelix,…

-**Minerals**: calcite, silica, alumina, titania, zeolite,…

-**Metals (alloys)**: bronze, steel,…

-**Allotropes**: Diamond, Graphite, monoclinic sulfur, ozone, …

-**General names**: table salt, vinegar,…

-**Other common names (mainly for small biochemicals)**: adenine, testosterone, mezerein, azalin B, mannitol, rosiglitazone, deferiprone, vitamin C,…

-**Amino acids**: standard names of amino acids (serine, asparagine,…) are considered as TRIVIAL, provided that there are not IUPAC-like indications (e.g. stereochemical terms).

NOTE: mentions of unique substances (not general family compounds) that are TRIVIAL and that are next to non-essential parts of the chemical entity or name modifiers (see M1, M4 and M9) should be assigned to the FAMILY class. This is especially important in patents as families of compounds are commonly written as the combination of a single "head" compound ("viagra") followed by the terms derivatives, analogues…

Viagra analogues	*tagged as FAMILY (FAMILY-TRIVIAL)*
Bupropion metabolites	*tagged as FAMILY*

But not:

Telmisartan medicinal composition	*tagged as TRIVIAL*
Benzbromarone tablets	*tagged as TRIVIAL*

CEM-5 (ABBREVIATION): this class covered the mentions of abbreviations and acronyms of chemical compounds and drugs. Only those abbreviations were annotated that could be clearly linked to chemical entities based on the annotators background knowledge or on descriptions provided in the article (ad-hoc abbreviations).

Annotate acronym, abbreviation and other definitions occurring before/after the chemical name separately:

[H]-8-OH-DPAT [8-hydroxy-2-(N,N-di-n-propylamino)tetralin]

2,4-dinitrophenyl)sulfenyl (DNPS)

Gamma-aminobutyric acid (GABA)

Where:

[3H]-8-OH-DPAT	*Systematic (systematic+formula +abbreviation)*
5-HT	*Systematic (systematic+abbreviation)*
8-hydroxy-2-(N,N-di-n-propylamino)tetralin	*Systematic*
(2,4-dinitrophenyl)sulfenyl	*Systematic*
DNPS	*Abbreviation*
Gamma-aminobutyric acid	*Systematic*
GABA	*Abbreviation*

Include acronym and abbreviation definitions that occur inside chemical names:

H-Lys-Trp(NPS)-OMe	*Formula (formula + abbreviation)*
proanthocyanidin (PAC) A2	*Trivial (trivial + abbreviation)*

Note: for non-common abbreviations → they should be tagged only if it is clearly an abbreviation of the chemical name of the compound:

hexahydro-1-nitroso-3,5-dinitro-1,3,5-triazine (MNX)

but not if they refer to a mixture of compounds or similar:

"…agrochemicals triadimefon and imazalil (MIX2) or triadimefon, imazalil, and the clinically used fluconazole (MIX3)…"

CEM-6 (FAMILY): this mention type includes well-defined chemical families that can be associated to some chemical structure. Pharmacological families were excluded from this class (refer to rule N2). This also **includes plural forms of systematic compound mentions that refer to a family of compounds**. In this case name-internal cues can be a useful help.

In this particular case the mentions of type FAMILY involve the following sub-categories as follows:

CEM 6.1 FAMILY-SYSTEMATIC CEM of type FAMILY that follows the systematic or semi-systematic nomenclature guidelines. Mainly plurals of IUPAC names:

Quinolines
Diacylglycerol ether
Pyridine compounds

As well as the terms referring to general chemical groups (aldehyde, hydroxide, amino acid,…). In case of doubt when the chemical entity may refer to either a single compound or a family of compounds (e.g. "urea"), the context should be considered to disambiguate.

As already mentioned (under SYSTEMATIC CLASS), mentions of unique substances (not general family compounds) that are IUPAC or IUPAC-like next to non-essential parts of the chemical entity or name modifiers should be assigned to the FAMILY class.

CEM 6.2 FAMILY-FORMULA CEM of type FAMILY that corresponds to a chemical formula (described in more detail in class FORMULA)
If the formula encompasses > 1 compound:

C-S-C bonds *Information on bonds/bridges (structural classes)*
ROH
CH stretching modes of DNP films
C-Cl bonds
C-H stretching
H-CO distance
(Asp3Phe1)2
dimer without specification on the specific connection
(Asp3Phe1)n
as n can be 1, 2… and includes peptides < 15aa, this CEM should be labeled as FAMILY (FAMILY-FORMULA)

CxF2x+1CH2C(O)H (x = 1,6)
CxF2x+1CH2C(O)OONO2 (x = 1,6)
Ir(x)Sn(1-x)O2

C10-C16 derivatives
C10-C16 alkyl esters
C-16/C-26 ester
-NH2- *note that both hyphens should be annotated*
But do not tag "=" when meaning equal (R = O)
And do not tag other symbols (e.g. arrows) when they are internal conventions found in patent documents
←O-

Note: CEM's bonded by non-covalent bonds should be annotated as MULTIPLE class because they are considered as two separate interacting CEMS.

Ar···propargyl alcohol *as MULTIPLE*
N-H···F-C *as MULTIPLE*

Note. Generic nomenclature is accepted within formulae only if the formula has more than 1 character. This is especially important in many patents.

MCl_2 where M is any metal	
[M-H](-)	
[M+Na](+)	
ROH stands for alcohols	
M = Cu, Ag	*M alone is not labeled*
R = amides, amines…	*R alone is not labeled*
X = any halogen	*X alone is not labeled*

CEM 6.3 FAMILY-TRIVIAL CEM of type FAMILY that corresponds to a trivial name (described in more detail in class TRIVIAL structural class names)

Terpenoids
Sugars
Wittig Reagent
Lewis Acid
Triacylglycerol

As already described (under TRIVIAL class), mentions of unique substances (not general family compounds) that are TRIVIAL and that are next to non-essential parts of the chemical entity or name modifiers (see M1, M4 and M9) should be assigned to the FAMILY class. This is especially important in patents as families of compounds are commonly written as the combination of a single "head" compound ("viagra") followed by the terms derivatives, analogues…

CEM 6.4 FAMILY-ABBREVIATION CEM of type FAMILY that corresponds to an acronym or abbreviation (described in more detail in class ABBREVIATION). Also, plural forms of abbreviations are to be considered of type FAMILY:

FAs = fatty acids	*as FAMILY (FAMILY-ABBREVIATION)*
ACs= anthocyanins	*as FAMILY (FAMILY-ABBREVIATION)*

CEM 6.5 FAMILY-FAMILY → other family names that do not match any of the other previous four classes. Are of the type family but one cannot clearly assign them to a more specific sub-class.

For example, adjectives in M4:

Pyrazolic compounds

NOTE: Synthetic polymers consisting of an undefined number of monomers (polyamide, polyvinylidene fluoride, PEG…) will be considered as FAMILY class members, independently on the rest of notation. The reason for this that, although in some cases the molecular weight is known, the molecular formula is general (e.g. they can be lineal or branched).

polyethylene glycol 6000	*as FAMILY (FAMILY-SYSTEMATIC)*
Brij30	*as FAMILY (FAMILY-TRIVIAL)*
Brij35	*as FAMILY (FAMILY-TRIVIAL)*
Gelucire 44/14	*as FAMILY (FAMILY-TRIVIAL)*
F127	*as FAMILY (FAMILY-TRIVIAL)*
F68	*as FAMILY (FAMILY-TRIVIAL)*
poloxamer 188	*as FAMILY (FAMILY-TRIVIAL)*
PEG	*as FAMILY (FAMILY-ABBREVIATION)*
PFS	*as FAMILY (FAMILY-ABBREVIATION)*
PBDEs	*as FAMILY (FAMILY-ABBREVIATION)*
Polybrominated diphenyl ethers	*as FAMILY (FAMILY-SYSTEMATIC)*
OH-PBDEs	*as FAMILY (FAMILY-FORMULA)*
PBDE-99	*as FAMILY (FAMILY-ABBREVIATION)*
PCB	*as FAMILY (FAMILY-ABBREVIATION)*
polychlorinated biphenyls	*as FAMILY (FAMILY-ABBREVIATION)*

CEM-7 (MULTIPLE): this class addressed mentions that did correspond to chemicals that are not described in a continuous string of characters. This is often the case of mentions of multiple chemicals joined by coordinated clauses or enumerations of chemical names (often used to avoid redundancies). Also parts of names divided by long text passages fall into this class. The dependencies of the partial chemical compound mentions are not captured in this version of the task. Such MULTIPLE mentions could be decomposed later defining the dependencies, chaining rules or alternative allowed mentions in a second step if needed. They are only annotated if the corresponding joined mention (integrated form) would be one of the other chemical entity mentions defined for this task.

7-[3-(fluoromethyl)piperazinyl]- and –(fluorohomopiperazinyl)quinolone compounds

thieno2,3-d and thieno3,2-d fused oxazin-4-ones

4-(3-chloro-4-hydroxyphenyl)- and 4-(4-chloro-3-hydroxyphenyl)-1,2,3,4-tetrahydroisoquinolines

phenylsulfenyl or acyclic sulfenyl substituted dipeptides

Hydroxy- and amino-substituted piperidinecarboxylic acids

Scrophularianines A-C

Note1: if there are terms inside the sentence that do not form part of the chemical name → they should not be tagged. Therefore, the potentially multiple entity will be splitted:

elaidic (t-C18:1 delta9) and palmitic acid	*two different entities*
methyl derivatives and ethyl carbonates	*two different entities*

N-Substituted and unsubstituted 4-chlorobenzene- and 4-nitrobenzenesulfonamides *unsubstituted adds no positive chemical information and it should not be tagged. Then, N-substituted is outside the MULTIPLE CEM and only N should be labelled as a separate CEM of type FORMULA.*

As already mentioned, CEM's bonded by non-covalent bonds should be annotated as MULTIPLE class because they are considered as two separate interacting CEMS.

Ar···propargyl alcohol	*as MULTIPLE*
N-H···F-C	*as MULTIPLE*

Note: in the case of anaphors within MULTIPLE mentions: the anaphor mention should be kept (although it is a nested mention):

schisarisanlactones A (1) and B (2)
reinocarnoside A (1), B (2) and C (3)

Note on the context: on how to deal with the context. In the case of specific, isolated CEMs that, when isolated correspond to a specific chemical entity but that in the context refer to a class of compounds → this CEM should be assigned to its family general class. Example:

In general the synthetic route involved the coupling of diethyl N-[2-fluoro-4-(prop-2-ynylamino)benzoyl]-L-glutamate with the ***appropriate*** *6-(bromomethyl)quinazoline followed by deprotection with mild alkali.*

6-(bromomethyl)quinazoline →	*should be tagged as FAMILY*

Ortography/Grammar Rules

O1 Other languages

Names in other languages than English should be annotated regardless the language according to the general annotation rules and CEM classes.

(9E)-9-Octadecensäure	*German*
9 trans - ácido octadecanoico	*Spanish*
9-octadecenoic acid, (9E)-Acide (9E)-9-octadécénoïque	*French*

O2 Mis-spellings & conversion errors

Mentions of chemicals (as long as they follow some of the other mention rules) that are misspelled should be tagged. This also includes mentions suffering from automatic conversion errors generated by text conversion programs.

ch1oro	*where 1 is "one" not "l"*
1. 1 equiv. Br2in dioxane, …	*where it should be "Br2 in dioxane"*

When manually annotating, the qualifier "CEM_TYPO" should be added to these mentions under the comments field.

O3 "A B" wrong space

White space-separated words that should properly be a single word → should be marked up as single entity.

… the acetoxy ethyl group was …

When manually annotating, the qualifier "CEM_TYPO" should be added to these mentions under the comments field.

O4 Chemicals named after people

Mentions of chemicals named after people should be tagged if they do refer to very clear chemical structures. These mentions correspond generally to "trivial" or "family" names widely used.

Tröger's base	*Trivial*
Schiff base	*Family-Trivial*
Grignard reagents	*Family-Trivial*

But this only applies for chemical entities (not chemical reactions):

Gewald thiophene synthesis	*only tag thiophene*

O5 Sentence boundary

Chemical entity mentions cannot span multiple sentences.

O6 Not short mentions

Do not tag acronyms that are of 1 letter in length. 1-letter code of amino acids/nucleotides or biochemical mutation mentions should be excluded. 1-letter code of chemical elements should be annotated (as FORMULA)

A T R Arg176Met	
1154C>T (A385V) and 1193T>C (M398T) in the coding exons	*untagged*
Pd/C	*Each of them tagged as FORMULA.*
N-terminal	*N (nitrogen should be tagged as CEM FORMULA)*
FA = Fatty Acid	*FA tagged as FAMILY (FAMILY-FORMULA)*
Arg-Lys-Phe (RKF)	*RKF tagged as FORMULA*
E(2)	*tagged as ABBREVIATION*

O7 Not flanking white space characters

Not tag white space characters flanking the CEM. Annotators should try to define the mentions precisely, and not include flanking whitespace or other spacing characters.

O8 Not Commas, full stops, brackets

Do not include as part of the CEM: off commas, full stops, brackets, and references to papers etc. that aren't a part of the name itself. Do include as part of CEM the square brackets around inorganic complexes and ionic liquids only if the bracket appears within the name.

[Co(CN)53I]

but:

$[Cu(H2O)6]^{2+}$

Acetate, bromine, the new compounds (aspirin and (carboxyalkyl)hydroxypyridinone)

Deferiprone (1,2-dimethyl-3-hydroxypyridin-4-one)

O9 Include prefixes for stereochemistry

Include in the CEM label prefixes that denote stereochemistry or regiochemistry of the compound.

cis-methanoglutamate
cis-platin
(S)-Alanine

(3RS,4SR)-4-acetamidopiperidine-3-carboxylic acid
cis-isomer 22 *nothing tagged (no anaphors o general terms)*

O10 Not Trademarks

Do not include trademark symbols as part of CEM

Aspirin®
Mesupron®

Unless the trademark symbol is in found within the CEM:

Tween® 80
Kollidon® VA64
Eudragit ® S100

O11 Not trailing hyphen/apostrophe

Do not tag trailing hyphens or the apostrophe-s in possessives. Exception: keep them in CAS names, keep them in case of FAMILY mentions.

Methyl-group
Kainite-preferring subunits GluR6 (*GluR6 is a protein receptor*)
Chloroform-induced ventricular tachycardia
Benzoic acid, 4-[[6-[[3'-(aminomethyl)[1,1'-biphenyl]-3-yl]oxy]-3,5-difluoro-2-pyridinyl]oxy]-
Benzene's activity

O12 Do not break up words to get at the CEM inside

Methylating *Not to be tagged (chemical reaction)*
Dienophile *Not to be tagged (reactivity role)*
Carbonium *To be tagged as ion (CEM), but not decomposed*
Acetyltransferase *Not to be tagged (enzyme)*
exo-**ATP**-site-directed reagents *ATP Not to be tagged inside the word*
m**Glu**R1alpha, m**Glu**R2 *Glu not to be tagged inside the receptors*

but:

ATP-site-directed inactivations
non-N-methyl-d-aspartate(non-NMDA) glutamate (Glu)

O13 Numbers in formula and numbers as part of the name

Include numbers on the front of formulae that indicate stoichiometry.

C6H8O3.2H2O *FORMULA*
$2H_2 + O_2 \rightarrow 2H_2O$ *FORMULA*

Include numbers that specify positions of a molecule only if they are part of the name:

C-2 carbon — *only carbon is annotated*
C-2 and C-3 positions — *nothing is annotated*
N-1 position "standard" substitution — *nothing is annotated*
…possessing a [4-hydroxy-3-(hydroxymethyl)-1-butyl] substituent at **N-1** exhibited an activity…

Ser473 — *only Ser is annotated*
Thr-384 — *only Thr is annotated*

If the positions identify general positions in compounds → these general positions should also be annotated

4-bromo derivative — *tag the 4- position*
5-vinyl substituent
5-[2-(1-azirinyl)]uracil analogues
5-vinyluracils
5-vinyl substituent of the respective 5-vinyluracils
2'-fluoro analogues
N-methyl derivative
5-[2-(1-azirinyl)]uracil analogues
with 5--19 spacer atoms between N6 or C-8 and iodine have been evaluated
do not tag the N6 and the C-8 positions

and if the position is within the CEM, label it:

brominated C(17) acetylenic acid

This rule on general positions applies for both numeric and string-defined (ortho, meta, para, o-…) positions in the molecule:

o-nitrophenyl-modified analogues

O14 State/charge/surface symbols

Include in the CEM oxidation state symbols, charge symbols, state symbols and surface symbols that occur on the end of names

Cu^{2+}
Cu(II)
CuSO4(*aq*)
Au(111) surface
(14)C isotope

MULTIWORDS: SINGLE ENTITIES vs MULTIPLE ENTITIES

M1 The longest CEM should always be tagged, but only including those words that are actually part of the chemical name. Non-essential parts of the chemical entity and name modifiers should NOT be tagged:

nitrogen gas
gold nanoparticules
methyl group
phenyl ring
caffeine analogue
carbon atom
cocaine addiction
Krebs citric acid cycle
Pyridine derivatives
Perovskite structure
Hydroxy-terminated conjugated polymer

but **substituted modifier should be tagged if inside a chemical entity** (meaning R – general formula assigned as FAMILY):

N-**substituted**-2-alkyl-3-hydroxy-4(1H)-pyridinones
chloro-**substituted** phenyls
6-fluoro-7-**substituted**-1,4-dihydro-4-oxoquinoline-3-carboxylic acids
2,4-diamino-5-(2',5'-**substituted** benzyl)pyrimidines
N-methyl-**substituted** sulfonamides

And similar modifiers as “substituted”:

(14)C-**labelled** lesogaberan	SYSTEMATIC
(13)C7-**labeled** iodoacetanilide	SYSTEMATIC
N(3)-**functionalized** xanthine	FAMILY
N-**interlinked** imipramines	FAMILY
pyridine-N-heterocyclic carbene-**based** palladacycles	FAMILY
N-**containing** polyketide	FAMILY
gadolinium-**embedded** iron oxide	SYSTEMATIC
hydroxy-**terminated** polyethylene glycol	FAMILY
S-**containing** arsenic metabolites	FAMILY
Paclitaxel-**conjugated** PAMAM dendrimers	FAMILY
pyrazole **amalgated** flavones	FAMILY
ent-3,4-seco-labdane-**type** diterpenes	FAMILY
deuterium-**enriched** ixabepilone	FAMILY
tricyclic nitrogen **containing** compounds	FAMILY
5-carbon-**linked**	FAMILY

but not if the word substituted (or similar words) do not provide specific information on the substitution (i.e., “isolated” words):

disubstituted naphtalenes
substituted 1,4-dihydronapthoquinones, hydroindoloquinones

amide alkyl substituents
14-substituted derivatives of carminomycinone
5-substituted acyclic pyrimidine nucleosides
A substituted or unsubstituted amino group
conjugated linoleic acid
N-Substituted and unsubstituted **4-chlorobenzene- and 4-nitrobenzenesulfonamides**

Note: the concept behind this annotation is that no attempt is made to establish meaningful correlations between different clearly separated CEMS (with non-annotated general words in-between) where the first CEM adds chemical information (typically substitution pattern) to the second CEM (however, this will be considered in future Biocreative tasks). If this were to be annotated, many clearly separated CEMs would collapse into a unique CEM that could not be translated into a chemical structure, thereby losing focus on the purpose of the annotation:

3-decladinosyl derivatives of 9-de-oxo-9a-aza9a-homoerythromycin A 9a,11-cyclic carbamate

Pyridine ring containing oxazilidinone

M2 Conflictive words: CEM or Modifiers? "Acid" "Base" "Salt" "Metal" "Radical" "Cation" "Anion" "Ion" "Dimer" (and similar to dimer)

Do only mark these words if they are part of a longer specific chemical name or if they refer to explicit classes of compounds (e.g. transition metal). Alone, these words should not be tagged (except for the case of the word "salt" meaning "sodium chloride").

Strong acid
Organic acid
Organic amines
lysergic acid
carboxylic acid
table salt
organic salt
citric acid trisodium salt
transition metal
metal oxide
heavy metal
the sodium salt
in treatment with aqueous alkali or acid *do not tag alkali / acid*
Aryl Diazonium Salts
sodium ion
hydroxyl radical
ammonium cation
thiol dimers
acetal trioxane dimer esters
p-phenylenevinylene trimers

M3 Adjectives with valid CEMs

Adjectives are only to be annotated if i) precede/follow a valid chemical entity and ii) add more precise structural information to this chemical entity. The whole concept (adjective + chemical noun) should be tagged as a unique chemical entity assignable to the chemical class of the chemical entity alone. This is independent on the origin of the root name of the adjective (i.e. systematic names or common names: pyrazolic vs nicotinic) and on the adjective ending ("-ed", "-ing","-olic").

polychlorinated biphenyl	
disubstituted naphtalenes	
acetylated phenoles	
dry ether	
ethanolic KOH	
allylic alcohol	
colloidal silver	
dry ice	*which is CO_2, not H_2O*
fuming sulphuric acid	*which is H2S2O7,not H2SO4*
warm HCl	
aqueous sodium carbonate	
molecular nitrogen	
primary alcohols	*specifies the precise type of alcohols*
secondary hydroxy groups	*specifies the precise type of hydroxyl groups*
stainless steel	
tertiary 2-(3-hydroxyphenyl)-2-phenethylamine	
(acetylated) glucuronide	
(saturated) hydrocarbons	
protonated nitrous oxide	

but not terms like "lower" that are very vague. These terms are typically found in patents and its exact meaning depends on each patent definition.

lower alkyl chain
linear or branched (C1-C6) alkyl

M4 Adjectives with general classes

Adjectives are only to be annotated if i) precede/follow a general compound class (compound(s), hit, analogue(s), derivative(s), series(s)…) and ii) add more precise structural information to this chemical entity (chemical class). Typically, these adjectives end as "-oic", "-oid", "-al", "-ois".

In contrast to M3, here only the adjective is tagged as a chemical noun of type FAMILY-FAMILY:

Pirazolic compounds	FAMILY
Terpenoids analogues	FAMILY
Aromatic organic molecules	FAMILY
Aromatic ring	FAMILY

Aromatic spacer	FAMILY
Aromatic group	FAMILY

But not if they still result in very wide compound families (commonly, adjectives finished in –ed correspond add less specific (R-group related) information than the others (-oic adjectives):

Methoxylated analogues	*nothing is tagged*
Fluorinated compounds	*nothing is tagged*

But not when found in different contexts:

glycemic control	*nothing tagged*
noradrenergic areas	*nothing tagged*

M5 Negative adjectives

"Negative" concepts that discard specific chemical structures but that do not explicit define a chemical structure should not be tagged.

2-desamino, 2-desamino-2-hydroxymethyl, and 2-desamino-2-methoxy analogues

desamino meaning "replace the amino group by hydrogen"

Similarly, the prefix *non-* should not be included:

non-steroidal	*tag only steroidal*
non-fluorinated parent compounds	*do not tag fluorinated as stated in M4.*

But it the term is to be tagged → then tag the corresponding adjective:

non fluorinated quinazolines	*tag the adjective*
non-fluorinated quinazolines	*tag the adjective*

M6 Enumerations and list of compounds vs multiple entities:

If full names are enumerated, tag separately each individual CEM:

citric acid and acetic acid
lithium carbonate, sodium carbonate
hexane-ethyl acetate, pyrane, aspirin/ibuprofen
aspirin, sugar, 4-methoxy phenol, and R-OH

If chemicals or class names of compounds are not described in a continuous string of characters→ tag the whole string (including words such as “and”, “or” and commas) as a single entity of class type multiple. Avoid the generation of “half truths”.

citric and acetic acid
lithium, sodium and potassium carbonate
pyrimidine derivatives and pyridine analogues
(as “pyridimide derivatives” is not a CM)
di(lower alkyl)amino
as lower is not tagged, tag separate CEMs
mono- and four new dimeric alkenylphenols
(dimeric adds meaning to the CEM. If it was “mono- and dimeric alkenylphenols” the whole CEM could be labelled as MULTIPLE. However, to avoid inclusion of non-CEM words (e.g. four), it is splitted and mono-, alone, is not meaningful)

When “>” acts as a separator (like a comma) of different CEMS, annotate each CEM as a single entity:

where R = methyl > ethyl > benzyl — tagged as different CEMs

H-bonds comes in the order of S1-H2···N2>N2-H2···S1>N3-H3B···O1 — tagged as different CEMS, as they have covalente bond, each of them is of type MULTIPLE

but if “>” acts as a qualifier of MULTIPLE entities, tag the whole CEM as MULTIPLE:

methyl < ethyl < propyl < butylparabens — tagged as a unique CEM MULTIPLE

M7 CEM Overlapping with Enzymes

Mentions of CEM that are part of mentions of enzymes should be tagged.

i. Two independent words where we only analyze the CEM:

K+ ATPase
Pyruvate kinase
phosphatidylinositol 3-kinase
metabotropic Glu receptors

ii. In the cases of hyphens we always split the words, and then they are independently analyzed:

Pyruvate-kinase
K+-ATPase

iii. Enzyme compound transformation "A B -ase", meaning "the -ase enzyme that catalyses the transformation of A to B", should be marked up as separate entities.

Squalene hopene cyclase (SHC) catalyzes the complex
Quinazoline antifolate thymidylate synthase inhibitors

But do not decompose CEMs:

Acetyltransferase

M8 CEM Overlapping with other non-chemical entities

Tag the corresponding chemical entity. For example, chemical formulae that appear inside mathematical formulae or equations (gradient, concentration):

^{1}H NMR
d[Na+]/dt = x
[caffeine]=10 mM

Although keep the brackets in the case of isotopes:

[3H]
[14C]

M9 CEM1 CEM2 → A single CEM or two CEMS?

If there are two continuous words of type CEM: “CEM1” and “CEM2”, each of which would individually be of class CEM:

- if they denote a single entity → label as a unique single CEM
- if they denote different chemical entities→ label as independent CEM’s

Use of adenine nucleotide derivatives → conceptually are a single entity (tagged as trivial)

NOTE!!! This criterion is not in agreement with rules defined by Corbett et al.2007, as we found that the strict classification of these rules (with interpretation) would be really time expensive and a potential source of discordance if no extra careful reading…

• **Generic terms that mirror IUPAC formation → a single entity**

Alkyl acetates
Isopropyl halides

• **Complexes and host-guest compounds defined by two continuous words → a single entit**y

Cu2+·2H20

- **Mixtures defined as "CEM1/CEM2" or "CEM1-CEM2"→ separate entities**

hexane-ethyl acetate *hexane =CEM1 and ethyl acetate = CEM2*
Pd/C preparation *Pd = CEM1 and C = CEM2*
isosteric benzene-thiophene replacement

but not all cases separated by '-' do correspond to mixtures. If the term refers to a single chemically meaningful compound or an aduct, it should be tagged as a single entity:

Terbinafine-HCl
TBF-HCl
Pt(II)-tetraphenyl-tetrabenzoporphyrin
quaternary ammonium salt-taxol

- **"the CEM2 that is part of the CEM1"→ a single entity**

carbonyl carbon
acetoxy methyl signal
acetoxy methyl group

- **"the CEM1 that is a CEM2" or " the CEM2 that is an CEM1"→ a single entity**

S-propionylthiolactyl-D-Glu-L-Lys thioester → *Difficult to differentiate that "thioester" is already implicitly mentioned in the previous CEM; by default, from practical perspective, we will annotate as unique CEM.*

terpenoid limonene → *We will separate them; since in this case "terpenoid" is an adjective and does NOT provide additional structural information to its corresponding name (as explained above). Therefore, in this case terpenoid will not be annotated*

pyrimidine nucleosides → tagged together as a single entity

- **"the CEM2 that contains an CEM1 group/moiety" → single entity**

Methyl ether
Tripeptide thioester

- **Terms ending in "glycoside"→ single entity**

Limonoid glycosides
Nominilic acid glycoside

Note: all these examples apply for the case of mentions next to each other. If the words are separated by other words, annotate them separately.