CHEMDNER data preparation and annotation manual

Version 2.0 (31th July 2013)

This document describes the data selection criteria and annotation guidelines used for the construction of the CHEMDNER task corpora. The annotation guidelines will be 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 reslease of the first sample set prepared for the CHEMDNER 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 the mention of chemical entities in text following a set of rules specified below. The text to be labeled consists mainly in PubMed abstracts (title and abstract text) in the first round of annotation followed by the annotation of a smaller set of full text scientific articles and patent abstracts.
When possible, 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 MyMiner and AnnotateIt annotation interfaces 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 LUKBXSAWLPMMSZ-OWOJBTEDSA-N
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 (this fine grained sub-annotation will only be done initially for a subset of the data collection).	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 task. For each CEM a short description and illustrative example cases are provided.

1. CHEMDNER chemical entities

The focus for defining the chemical entities annotated for the CHEMDNER task was primarily to capture those types of mentions that are of practical relevance. Therefore the covered chemical entities had to represent those kinds of mentions that can be exploited for linking articles to chemical structure information.

The annotation carried out for the CHEMDNER 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 task was that they could be associated to chemical structure information to at least a certain degree of reliability. This implied that very 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 yellow.

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.

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…*

Glucocorticoid *structurally similar, but without a strict group definition*

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, Aminoacids, 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, …

P3. Small Biochemicals

-Sacharids: 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 aminoacids was chosen as cut-off. Thus, peptides with less than 15 aminoacids should be tagged as CEM (both, cyclic and non-cyclic peptides).

Glutathione (*trimer*)
Cyclosporin A *11 aminoacids*)
Degarelix
Gonadotropin-releasing hormone (GnRH) *with 10 aminoacids*
Luteinizing-hormone-releasing hormone (LHRH) *same case as for GnRH*
Azaline B *small peptide with < 15 aminoacids*
Angiotensin *<10 aminoacids*

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 aminoacids), so it should not be tagged. In the same way, chemically modified proteins with > 15 aminoacids should not be tagged.

Luteinizing hormone (LH) *untagged because it has 92 aminoacids*

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

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

-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
Polyvinyl chloride (PVC)
Polyamides

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

-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:

Deshidrogentation
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,…

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

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

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

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

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

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

N3. Very nonspecific structural concepts:

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

> The stereoisomer 6, but not 7, activated cloned carminomycinone-aglycone (II) of carminomicin *not tagged*

> Refer to M2 for the special case of conflictive words: acid, salt, metal

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

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*

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 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 aminoacids*), annexin, atrial natriuretic peptide (*28 aminoacids*), peptide,

N6. General vague compositions

> Pigments with a relatively varying mixture: melanin

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+)

Lead → as it is a very common word in many chemical texts, meaning the "main" candidate compound from a chemical series or the verb "guide". As the expected chance of meaning the chemical element "lead" is much lower, we agreed in not including this word.

Gold Same as for lead

Note: In opposition to "*lead*" → the word "*iron*" should be tagged as within chemical texts it is much more probable to find this word referring to the chemical element than to the "cleaning" activity.

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 raking 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
Polystyrene

Here we also include the mention 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 M7):

2,3-Dihydrobenzofuran analogues

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

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*
D-Ala-D-Ala *3-letter codes of small peptides*
Glu-Cys-Gly *3-letter codes of small peptides*
GlcNAc *Oligosaccharides nomenclature: formula with abbreviation*
Asp-Glu-Fmoc *Formula (formula with abbreviation)*
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
t-BuOK

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, pyruvate kinase, xanthine oxidase, deferiprone,…

Note that the name of the amino acids (serine, asparagine,…) is IUPAC, so they should be labeled as SYSTEMATIC.

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	*Formula (formula + 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)*

CEM-6 (FAMILY)*:* this mention type included 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 included **plural forms** of systematic compound mentions that refer to a family of compounds. In this case name-internal cues can be a useful help. Initially the organizers planned to remove this class distributing the associated entities to other mention types. We finally decided to keep this as a separate CEM class because it involved chemical structural information and in some cases is of practical relevance.

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

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.

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

Note. Generic nomenclature is accepted within formulae only if the formula has more than 1 character:

MCl_2 where M is any metal
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

Synthetic polymers consisting of an undefined number of monomers (polyamide, polyvinylidene fluoride, PEG…) will be considered as FAMILY class members.

CEM 6.4 FAMILY-ABBREVIATION CEM of type FAMILY that corresponds to an acronym or abbreviation (described in more detail in class 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

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 antibacterials

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

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*

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.*

Note2: 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 non-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"*

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 …

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 or 2 letters in length. 1-letter code of aminoacids/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 *these are tagged because they are of CEM FORMULA*

N-terminal N (nitrogen should be tagged as CEM FORMULA)

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®

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
Acyl-enzyme inhibitors

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* n*ot to be tagged inside the receptors*

but:

ATP-site-directed inactivations
anti-dopamine beta-hydroxylase
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

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:

sodium ion
hydroxyl radical
nitrogen gas
gold nanoparticules
methyl group
phenyl ring
caffeine analogue
carbon atom
cocaine addiction
Krebs citric acid cycle
Pyridine derivatives
Perovskite structure

but **substituted modifier should be tagged if inside a chemical entity** (meaning R):

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

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
N-Substituted and unsubstituted **4-chlorobenzene- and 4-nitrobenzenesulfonamides**

M2 Conflictive words: CEM or Modifiers? "Acid" "Base" "Salt" "Metal"

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
lysergic acid
carboxylic acid

table salt *incluso de esta se podría hacer una exception como water*
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*

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

ionotropic glutamate receptors *do not tag "ionotropic"*

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-Family
Terpenoids analogues	Family-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)

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

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

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 entity

Cu2+·2H20
Hydroxypropil-beta-cyclodextrin-itraconazole

• 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

• "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.

A complex of hydroxypropyl-bet-cyclodextrin and itraconazole