Appendix B

Manual for ESMILES Notation

This manual gives description about the input needed to generate templates and for prediction. It discusses the format by which the educts and products should be provided so that the correct templates is generated and also the format by which educts should be given as input for proper prediction of products. In this manual the format described is in Extended simplified molecular input line-entry system or ESMILES notation. The details of ESMILES notation is described below with examples.

1 ESMILES notation of Chemical compounds and chemical bonds

Every compound in ESMILES notation is enclosed under ("{""}") curly braces. This is done to separate each compound from one another for processing. The other parts of ESMILES notation is discussed below

1.1 Single Bond

The single bond is represented by single dashed line (-). Two atoms containing '-' symbol between them denotes they are connected by a single bond.

Let "A" and "B" denotes two elements.

Generalised form : A-B

ESMILES form: {A-B}

Examples

1. H_2 (Hydrogen gas) is represented by {H-H} in ESMILES notation.
2. HCl (Hydrochloric acid) is represented by {H-Cl}.

1.2 Double Bond

The double bond is represented by equals to sign (=). Two atoms containing '=' symbol between them denotes they are connected by double bond.

Let "A" and "B" denotes two elements.

Generalised form : A=B

ESMILES form: {A=B}

Examples

1. O_2 (Oxygen gas) is represented by {O=O} in ESMILES notation.
2. CO_2 (Carbon di oxide) is represented by {O=C=O}. in ESMILES notation.

1.3 Triple Bond

The triple bond is represented by hash symbol (#). Two atoms containing '#' symbol denoted they are connected by triple bond.

Let "A" and "B" denotes two elements.

Generalised form : A≡B

ESMILES form: {A#B}

Examples

1. CO (Carbon mono oxide) is represented by {C#O} in ESMILES notation.
2. HCN (Hydrogen Cyanide) is represented by {H-C#N}. in ESMILES notation.

1.4 Coordinate Bond

The coordinate bond is represented by tilde (∼) or tilde(∼) + '/' sign (∼/) sign depending upon the potion of donor atom and receiver atom.

Let "A" and "B" denotes two elements.

Generalised form : A->B

ESMILES form: {A∼B} or {B∼/A}

Examples

1. NH is represented by {N∼H} in ESMILES notation as N (nitrogen) is donor and H(hydrogen) is receiver atom, but the same compound can be represented by {H∼/N} when the positions of N and H changes.
2. NB is represented by {N∼B} in ESMILES notation as N (nitrogen) is donor and B(boron) is receiver atom, but the same compound can be represented by {B∼/N} when the positions of N and B changes.

1.5 Cations

Cations are represented in ESMILES notations by putting '+' sign along with the magnitude beside the atom.

Let "A" denote an element

Generalised form : A^{+2}

ESMILES form: {A+2}

Examples

1 H^{+} Hydrogen ion is represented as {H+1} in ESMILES notations.
2 Fe^{+2} Ferrous ion is represented as {Fe+2} in ESMILES notations.

1.6 Anions

Anions are represented in ESMILES notations by putting '-' sign along with the magnitude beside the atom.

Let "A" denote an element

Generalised form : A^{-2}

ESMILES form: {A-2}

Examples

Example :
1. Br^{-} (bromide ion) is represented as {Br-1} in ESMILES notations.
2. O^{2-} (oxide ion) is represented as {O-2} in ESMILES notations.

1.7 Free Radical

Free radical(s) is represented by a dot followed by a number representing number of unpaired valence electrons after the atom.

Let "A" denote an element

Generalised form : :A:

ESMILES form: {A.4}

Examples

1. Cl. (chlorine) is represented as {Cl.1}
2. :CO: (Carbon Mono Oxide) is represented as {C.2#O.2}.

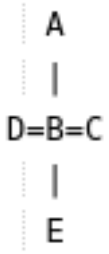
1.8 Branched Compounds

Branched compounds are represented by using parenthesis in ESMILES notation. When an atom has more than one compound or bond attached to

it, the compounds/bonds are separated by enclosing them inside parenthesis ("()").

Let "A", "B", "C", "D" and "E" denotes five elements

Generalised form :



ESMILES form: {B(=D)(=C)(-A)(-E)} or {D=B(=C)(-A)(-E)} or {A-B(=C)(=D)(-E)}

Examples

1. CO_2 (carbon di oxide) is represented as {C(=O)(=O)}
2. H_2SO_4 (sulphuric acid) is represented {S(=O)(=O)(-O-H)(-O-H)} , another representation of H_2SO_4 can be {H-O(-S(=O)(=O)(-O-H))}

1.9 Ionic Compounds

In ESMILES notations ionic compounds are enclosed inside third brackets('[]'), separately. The atoms under first pair of third brackets are cations and atoms under second pair of third brackets are anions. Both cations and anions are written side by side.

Let "A" and "B" denotes two elements.

Generalised form : A^+B^-

ESMILES form: {[A][B]}

Examples

1. NaCl (Sodium chloride) is represented as {[Na][Cl]}.
In here [Na] represents cation in the ionic compound and [Cl] represents anion in the ionic bond.
2. NH_4NO_3 (Ammonium Nitrate) is represented as
{[N(-H)(-H)(-H)(∼H)][N(=O)(-O)(-O)]}.
Here also [N(-H)(-H)(-H)(∼H)] represents cations in the ionic compound and [N(=O)(-O)(-O)] represents anions in the ionic compound.

1.10 Cyclic Compounds

The proposed ESMILES notations for cyclic compound is improved from SMILES notation by indicating the type of bond present between the two atoms along with the cycle number. The cycle number denotes which two atoms/ions forms bond in the cyclic compound whose bond is not shown in the ESMILES notation. The bond type denotes the bond between those atoms. The cycle number and bond type is enclosed in square brackets("[]") and written beside both the atom which is a part of the cycle.The bond type can be single('-'), double('='), triple('#') or coordinate('∼','∼/').

Let "A", "B", "C", "D" denotes 4 elements

Generalised form :

ESMILES form:
{A[1-]=B-D-C[1-]} or
{A[1=]-C=D-B[1=]}

Examples

1. C_6H_6 (benzene) can be represented as
{C[1=](-H)(-C(-H)(=C(-H)(-C(-H)(=C(-H)(-C[1=]-H)))))}.
Here there are two 'C[1=]' which represents bond number and bond type. it shows that the two carbon atoms connected by double bond.
2. HN_5 (pentazola) can be represented as {H-N[1-]-N=N-N=N[1-]},
Here 'N[1-]' represents both the nitrogen atoms are connected by single bond.

1.11 Coordination Complex Compounds

Coordination complex compounds are generally ionic compound, so they also enclosed under square brackets as done for ionic compounds. In ESMILES notation , the coordination centre is written first and then the ligands are written in parenthesis as denoted in branched compound.

Let "A", "B", "C", "D" denotes 4 elements

Generalised form : A^+A^+ $[C->B<-D]^{2-}$

ESMILES form: {[AA][B(<∼C)(<∼D)]}

Examples

1. $K_4[Fe(CN)_6]$ is represented as
{[KKKK][Fe(∼/C#N)(∼/C#N)(∼/C#N)(∼/C#N)(∼/C#N)(∼/C#N)]}.
Here [Fe(∼/C#N)(∼/C#N)(∼/C#N)(∼/C#N)(∼/C#N)(∼/C#N)] is the coordination complex compound, where 'Fe' is the coordination centre and 'C#N' is the ligand.
2. $[Co(NH_3)_6]Cl_3$ is represented as
{[Co(∼/N(-H)(-H)(-H))(∼/N(-H)(-H)(-H))(∼/N(-H)(-H)(-H))][ClClCl]}.
Here [Co(∼/N(-H)(-H)(-H))(∼/N(-H)(-H)(-H))(∼/N(-H)(-H)(-H))] is the coordinate compound with 'Co' as the coordination centre and N(-H)(-H)(-H)" as the ligand.

1.12 Addition Compounds

In ESMILES notations an addition compound is represented by an asterisk sign (*) placed in between the two compounds. The addition compounds are each written under first parenthesis.

Let "A", "B", "C", "D" denotes 4 elements

Generalised form : A-B.C-D

ESMILES form: {A-B*(C-D)}

Examples

1. $CuSO_4.5H_2O$ (Copper Sulphate penta hydrate) is represented in ESMILES as
[Cu][S(=O.4)(=O.4)(-O.6)(-O.6)]*(H-O.4-H)(H-O.4-H)(H-O.4-H)(H-O.4-H)(H-O.4-H)
2. $NaCl.H_2O$ (Sodium Chloride mono hydrate) is represented in ESMILES as [Na][Cl]*(H-O.4-H)